UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

LincRNA profile in clear cell Renal Cell Carcinoma using RNA-seq data

Ioana Posa

Dissertação

Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

2014

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

LincRNA profile in clear cell Renal Cell Carcinoma using RNA-seq data

Ioana Posa

Dissertação

Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

Dissertação orientada pela Prof. Doutoura Lisete Maria Ribeiro de Sousa
e pela Doutoura Ana Rita Fialho Grosso

2014

Agradecimentos

Em primeiro lugar gostaria de agradecer à Doutora Ana Rita Grosso pela constante ajuda, otimismo, confiança e entusiamo transmitido em alturas difíceis e sem esperança à vista.

Agradeço também ao Doutor e Principal investigador Sérgio Almeida por me ter recebido no seu grupo de investigação do Instituto de Medicina Molecular. Ao Doutor Nuno Barbosa Morais pelo seu entusiasmo transmitido e ajuda ao longo do desenvolvimento do trabalho. À Professora Doutoura Lisete Maria Ribeiro de Sousa pela sua prontidão em resolver problemas burocráticos e pelos seus *insights* estatísticos.

Ao suporte informático do Instituto de Medicina Molecular, sempre pronto a resolver os meus problemas computacionalmente existenciais.

À minha colega Joana Tavares pela luta e entreajuda constante que tivemos ao longo deste ano para percebermos o bem o mal que estávamos a fazer ao longo do desenvolvimento das várias etapas do trabalho.

Aos meus amigos, que sem vocês tinha-me perdido neste mundo digital e me trouxeram muitas vezes à realidade.

Por fim gostaria de agradecer e dedicar esta tese principalmente aos meus pais que sacrificaram muitos e bons anos de vida para eu poder ter as melhores oportunidades que alguém pode ter. Vocês são as pessoas mais importantes da minha vida e sem vocês nunca poderia ser a pessoa que sou hoje. Um eterno obrigada por tudo o que fizeram e sacrificaram por mim e espero que um dia mais tarde possa restituir um pouco da vossa felicidade.

Para os meus pais.

Resumo

O cancro renal ou carcinoma de células renais (*renal cell carcinoma* - RCC) é um grupo comum de doenças resistentes a quimioterapia. É um dos tipos de cancro mais letal no sistema urinário, sendo a taxa de sobrevivência para os pacientes com RCC metastático de menos de 10% após cinco anos de diagnóstico.

Com base nas suas características genéticas e histológicas, no seu fenótipo clínico e diferentes respostas à terapia, os RCCs podem ser subdivididos em vários tipos, sendo um dos mais comuns o de células claras RCC (*clear cell renal cell carcinoma* - ccRCC); correspondendo a mais de 80% dos casos de RCC.

Uma das características do ccRCC, bem como de outros tipos de cancro, é a metabolização da glucose através da glicólise seguido pela produção de lactato, processo primeiramente descrito por Warburg - " efeito de Warburg ". Este efeito ocorre em oposição à normal glicólise seguida de fosforilação oxidativa mitocondrial, a fim de produzir o adenosina trifosfato (ATP).

Esta característica deriva principalmente do gene von Hippel-Lindau (VHL) inactivo, contudo este apenas apresenta mutações que podem inactivar a sua função em apenas 52% das amostras de ccRCC. Esta mutação poderá não ser suficiente para explicar este carcinoma e que mais estudos são necessários a fim de entendê-la. Um papel importante neste cancro também tem sido atribuído à regulação epigenética, bem como a microRNAs desregulados.

Desde o início do século XXI, vários projectos a nível global têm permitido descartar a ideia de que o genoma humano é principalmente " lixo" e para isso também contribuiu o desenvolvimento de tecnologias de sequenciação de nova geração (*next generation sequencing* – NGS).

Algumas destas tecnologias são a Roche 454, Illumina / Solexa e tecnologias ABI-Solid que permitem sequenciar todo o genoma/ transcriptoma de uma só vez. De modo a ocorrer esta sequenciação é necessária uma fragmentação do material genético; uma reacção em cadeia da polimerase (*polimerase chain reaction* - PCR) em paralelo e determinação da sequência através de fluorescência. Nos últimos anos, surgiram tecnologias de sequenciação de terceira geração (como PacBio e Helicos) capazes de definir a sequência utilizando moléculas individuais de DNA, sem necessidade de reacções de PCR.

Actualmente, grande parte das tecnologias disponíveis apresentam várias vantagens

e limitações, sendo o mais importante na escolha de uma destas o equilíbrio entre os objectivos e o orçamento disponível.

Uma das técnicas que tira partido destas tecnologias é a sequenciação de RNA (RNA-seq). Esta técnica, utiliza tecnologias de sequenciação de nova geração, a fim de analisar todas as moléculas de ácido ribonucleico (*ribonucleic acid* - RNA) de uma ou mais células – transcriptoma. A análise deste tipo de dados permite fornecer informações a nível de sequência, bem como sobre níveis de transcrição, facilitando o desenvolvimento de novas terapêuticas e interpretação de dados experimentais.

Esta revolução tecnológica levou ao reconhecimento de que o transcriptoma não é apenas constituído por transcritos codificantes de proteínas, mas também por um elevado número de transcritos não codificantes. Transcritos estes que estão a ser gerados a partir de regiões que se acreditava ser "desertos". A transcrição generalizada das regiões não codificantes pode estar na origem de moléculas funcionais. Torna-se assim evidente que existe uma necessidade de ter em conta elementos não codificantes, ao serem realizados estudos de associação ao nível do genoma.

Os transcritos não codificantes (*non coding RNA* - ncRNA) estão associados a várias funções a nível celular e pedem ser divididos em várias categorias, de acordo com o seu tamanho e localização relativa a genes codificantes de proteína. Um dos grupos de ncRNA são os longos transcritos não codificantes intergénicos (*long non coding intergenic RNA – lincRNA*), que apresentam um tamanho superior a 200 nucleótidos e não apresentam nenhuma sobreposição com outros genes anotados.

Estes não apresentam nenhuma característica específica sendo que podem ser transcritos pela mesma maquinaria que permite a transcrição de genes codificantes de proteínas. Normalmente apresentam cerca de 2 a 3 exões e o nível de expressão é menos elevado que o dos genes codificantes. O papel biológico da maioria destes ainda é, em grande parte desconhecido, contudo alguns deles têm sido associados a vários tipos de cancro.

Apesar da quantidade de estudos feitos em ccRCC e da quantidade de mutações identificadas, ainda não é possível compreender este subtipo de carcinoma renal. Assim, decidiu-se explorar o perfil de expressão de lincRNAs em ccRCC e quantificar diferença na expressão destes, comparando amostras normais versus a amostras de tumor de 62 pacientes com ccRCC. Para isso, é necessário construir o transcriptoma base do ccRCC para a descoberta de potenciais novos lincRNAs; analisar a expressão diferencial de lincRNA e mostrar sua correlação com genes que codificam proteínas. Foi então utilizada uma análise computacional de dados de RNA-seq de 62 amostras de pacientes ccRCC (pares de amostra tumoral e normal).

Primeirament foi construido um catálogo com lincRNAs humanos, utilizadando anotações de lincRNA de várias bases de dados (Ensembl, Gencode, Vega, Lncipedia, UCSC, do Instituto Broad, Noncode e dados publicados por Zhipeng e Adelson). A falta de correspondência entre as diferentes bases de dados, aumentou o grau de complexidade do

processo, contudo no final foi obtido um catálogo de 38 134 lincRNAs humanos.

De seguida, foi reconstruído o transcriptoma do ccRCC para usar como base para nova descoberta de lincRNAs. A caracterização das 62 amostras de pacientes ccRCC (tumor e normal combinado) revelou 5549 potenciais novos linRNAs. A análise diferencial entre as amostras de cancro e tecido normal permitiu a identificação de 2129 genes diferencialmente expressos (entre os quais 239 lincRNA e 105 potencias novos lincRNAs).

Devido aos seus baixos níveis de expressão, para muitos dos lincRNAS o teste estatístico não foi sequer efectuado. Facto pelo qual, o último passo envolveu uma análise que tem em conta a relação entre os transcritos, independentemente da sua expressão diferencial. Foi realizada uma análise de correlação génica em rede (*gene correlation network analysis*), permitindo encontrar genes altamente correlacionados entre si e o tipo de amostra - tumor / normal. É de realçar o lincRNA PVT1, que foi previamente associado a outros tipos de cancro e tem uma elevada expressaão em amostras de tumor ccRCC. Pacientes com elevada expressaão relativa deste lincRNA nas amostras normais, têm uma probabilidade inferior de sobrevivência comparativamente aos que apresentam uma menor expressão relativa.

No final, esta análise permitiu a dar os primeiros na compreensão a importância dos lincRNAs no ccRCC.

Palavras chave: lincRNA, ccRCC, RNA-seq, diferencialmente expresso, FPKM, rede de correlação génica.

Abstract

Kidney cancer or renal cell carcinoma (RCC) is a common group of chemotherapy resistant diseases, and one of the most lethal type of cancer in the urinary system, being the survival rate for patients suffering from metastatic RCC is less than 10% survive five years subsequent to diagnosis.

Based on their genetic characteristics, histological features, clinical phenotype and different responses to therapy, RCCs can be subdivided in several subtypes, one of the most common being clear cell RCC (ccRCC) accounting for more than 80% of RCC cases. ccRCC is usually characterized with an inactive von Hippel–Lindau (VHL) gene, the VHL gene mutations that can inactivate were observed only in 52% samples, which may indicate that this mutation is not sufficient to explain this carcinoma and that more studies are necessary in order to understand it. An important role for epigenetic regulation has also been suggested for ccRCC, as well for deregulated microRNAs.

The development of next generation sequencing technologies (NGS) made possible for a bigger number of transcriptomes to be analysed. This allowed to acknowledge that a transcriptome is not only constitute by protein-coding transcripts but also by a high number of non-coding transcripts. This transcripts are being transcribed from regions previously thought to be "deserts". This widespread transcription of non-coding regions may be in the origin of functional molecules, making apparent that there is a need to take into account non-coding elements when genome wide association studies are done.

Non-coding RNA (ncRNA) are associated with plenty of functions and one group of ncRNA - long intergenic ncRNA, which have no overlap other annotated genes, have been associated with several other cancers.

Despite the amount of studies made in ccRCC and the amount of identified mutations it is still not possible to comprehend this subtype of renal carcinoma. Thus, we decided to explore the long intergenic non-coding RNA (lincRNA) profile in ccRCC and quantify difference in gene expression when comparing the normal versus the tumor samples.

For that is necessary to assemble the ccRCC transcriptome as base for potentially new lincRNA discovery, analyse differential lincRNA expression and show their correlation with protein coding genes.

In order to achieve that, a computational analysis of RNA-seq pair-end data of 62 ccRCC patient samples (tumor and matched normal) was used.

In order to accomplish these objectives, a human lincRNA catalog, with lincRNA annotations from several databases (Ensembl, Gencode, Vega, Lncipedia, UCSC, Broad Institute, Noncode and Zhipeng and Adelson published data) had to be constructed. The main preoccupation was to have the most complete tool/resource for assessing lincRNA expression. For that, 8 different databases with lincRNA annotations were merged in order to obtain a unified human catalogue of 38 134 lincRNAs.

To uncover the lincRNA profile in ccRCC, the transcriptome composition of 62 ccRCC patient samples (tumor and matched normal) was assessed. Available bioinformatic tools were used and made possible the identification of 5549 potentially new lincRNA and determine 2129 differentially expressed genes (239 lincRNA and 105 potentially new lincRNAs).

In order to proceed with an analysis that takes into account the relationship between the transcripts, independently of their differential expression, a weighted gene correlation network analysis followed. This analysis allowed to find highly co-expressed/correlated genes as well as genes highly correlated with sample type – tumor/normal sample, leading to uncover PVT1 lincRNA. This lincRNA was already associated with other cancers and has an expression highly upregulated in ccRCC tumor samples. Patients with relative high expression of this lincRNA in normal samples also show poor survival chances.

In the end, this analysis allowed to give the first steps in order to understand the lincRNAs importance in ccRCC.

Key words: lincRNA, ccRCC, RNA-seq, differentially expressed, FPKM, gene correlation network.

Contents

List of Tables

List of Figures

List of Abbreviations

ATP adenosine triphosphate

bp Base pair(s)

ccRCC clear cell Renal Cell Carcinoma

cDNA complementary DNA

DNA Deoxyribonucleic acid

FDR false discovery rate

FPKM fragments per kilobase of exon per million fragments

Glu1 glucose transporter 1

H3K36me3 trimethylation of lysine 36 in histone H3

H3K4me3 trimethylation of lysine 4 in histone H3

HIF-α hypoxia-inducible factor subunit α

HIF-β hypoxia-inducible factor subunit -β

HRE hypoxia response element

LDH lactate dehydrogenase

lincRNA long intergenic non-coding RNA

lncRNA long non-coding RNA

MCT4 monocarboxilate transporter 4

mRNA messenger RNA

mTOR mammalian target of rapamycin

NADPH nicotinamide adenine dinucleotide phosphate

ncRNA non-coding RNA

NGS Next-Generation Sequencing

nt nucleotide(s)

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

PDK-1 pyruvate dehydrogenase kinase 1

PPP pentose phosphate pathway

pre-mRNA pre-messenger RNA

RCC Renal Cell Carcinoma

RNA-seq RNA sequencing

RNA Ribonucleic acid

rRNA ribossomal RNA

TCGA The Cancer Genome Atlas

VEGF vascular endothelial growth factor

VHL von Hippel–Lindau tumor suppressor gene

WGCNA Weighted gene correlation network analysis

Chapter 1

Introduction

From the beginning of the XXI century projects like **ENCODE** (**Enc**yclopedia **O**f **D**NA **E**lements) (The ENCODE Project Consortium, 2004) allowed to dismiss the view that the human genome is mostly "junk DNA", hence 80% of the genome contains elements linked to biochemical functions (Ecker et al., 2012), bringing to light, among other elements, transcripts that have little protein-coding capacity (Gingeras, 2007) – non-coding ribonucleic acid (other than ribossomic RNA or transfer RNA), but that may have function. Together with this came the development of high throughput sequencing technologies, and the world entered in a new genomic era that brought huge impacts on genetic applications like metagenomics, comparative genomics, high throughput polymorphism detection, analysis of small RNAs, mutation screening, transcriptome profiling, methylation profiling, and chromatin remodeling (Yadav et al., 2014) .

1.1 Gene expression

According to HUGO's gene Nomenclature Committee (Wain et al., 2002), a gene is defined as a deoxyribonucleic acid (DNA) segment that contributes to phenotype/function and in the absence of demonstrated function it may be characterized by sequence, transcription or homology.

The information present in a **gene** is **transformed into a functional gene product**, being this a **protein or a functional ribonucleic acid (RNA)**, through ***gene expression***. If the functional gene product is a protein, DNA needs to be transcribed to a pre-messenger RNA (pre-mRNA) by RNA polymerase, processed into a mature messenger RNA (mRNA), translated into a polypeptide in the cytoplasm, folded into proper three-dimensional structure. On the contrary, if the final product is a functional RNA, DNA is transcribed into a non-coding RNA (ncRNA) by RNA polymerase and processed into a functional RNA.

In the human genome, gene transcription can be processed by three different nuclear RNA polymerases and a forth that is present in the mitochondria. The nuclear poly-

merases have distinct functions and use different promoters. Protein coding genes are transcribed by RNA polymerase II whereas genes that result into a functional RNA may be transcribed by RNA polymerase I, II or III, depending on RNA type (Strachan and Read, 2010).

Using RNA-sequencing (RNA-seq) technology, Djebali and colleagues (Djebali et al., 2012) report evidence that **75% of the human genome can be transcribed at some point**, as well an increased overlap of genic regions due to transcripts being synthesized from both strands, forcing a review of the actual gene concept and the minimum unit of heredity, and bringing more plausibility to the concept that it is possible for a gene to transcribe for multiple different transcripts (coding and non-coding) and have multiple regulatory regions (Gingeras, 2007).

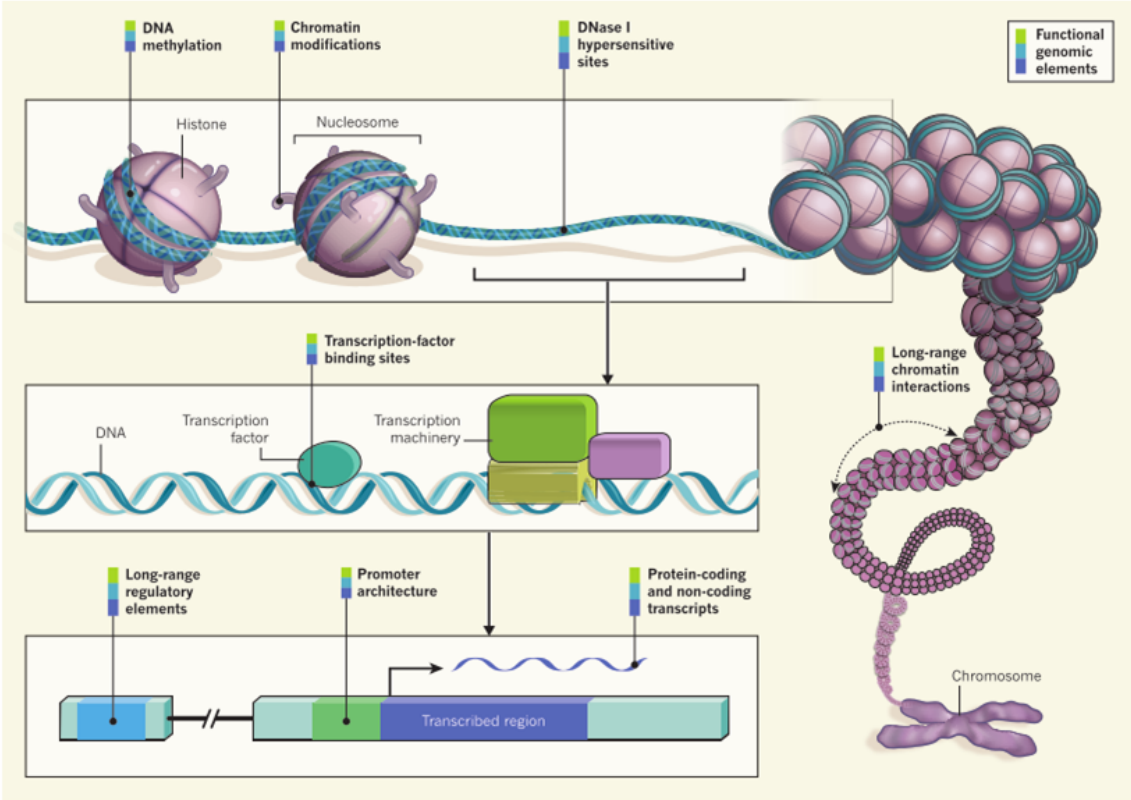


Figure 1.1: Global view of what happens in order to gene transcription occur and several elements that can intervene in this process. Here is possible to observe that in order for a gene to be transcribed, is necessary a decondensation of the chromosome (through the action of DNA demethylations and histone modification enzymes) in order for the DNA strand to be accessible for the transcriptional machinery. Several others factors, as chromatin interactions, transcription factors binding sites or long range regulatory elements, among others, can also induce/repress the transcription of a gene. Image adapted from Ecker et al. 2012

In this new genomic era it becomes clearer that gene transcription is a well coordinate choreography between DNA structure, sequence and regulatory elements. Chemical modification of histones, DNA methylation state can influence the transcription of DNA

(Ecker et al., 2012) as well as other regulatory elements (desoxirribonuclease I, binding sites, transcription factor binding site, long range regulatory binding sites, DNA binding proteins, RNA molecules, etc). In **Figure 1.1** it is possible to observe a general view of what happens in order for a gene to be transcribed into a protein-coding or a non-coding transcript.

1.2 High throughput sequencing

With the DNA double helix structure identification by Watson and Crick in 1953, the world started a new journey in order to discover all genome's particularities. Sanger *et all* (Sanger et al., 1977), developed the di-deoxi chain termination method in 1977, which allowed to determine DNA sequence of a certain region; method broadly used by scientists until a decade ago. Only in the 2000's started to appear technologies capable to sequence all the genome at once. Nowadays, a lot of technologies are available, presenting all advantages and limitations; being the most important when choosing one of them, the balance between goals and the available money. In **Figure 1.2** it is possible to observe the evolution of DNA sequencing technologies along the years.

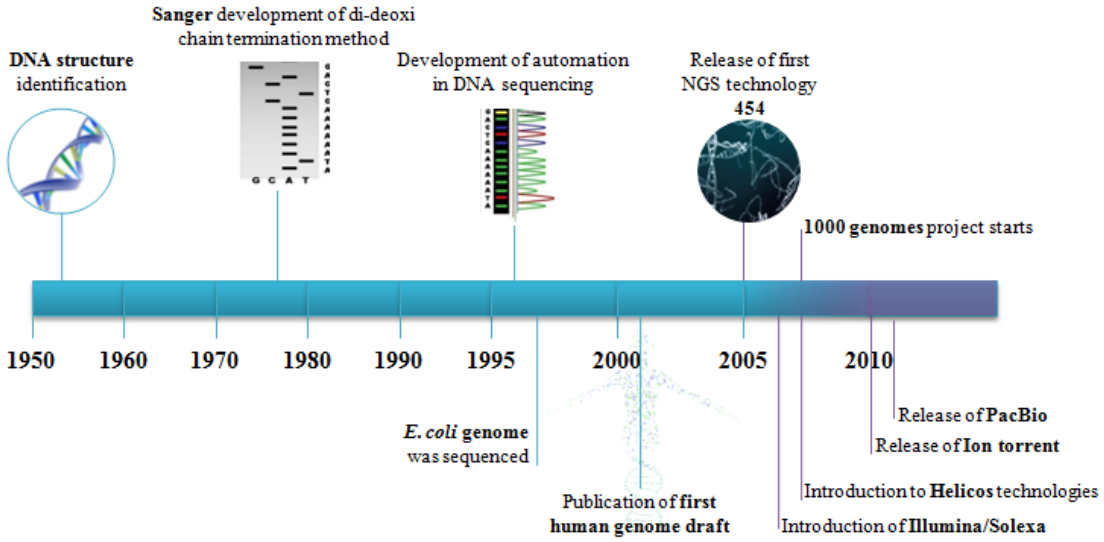


Figure 1.2: Evolution of DNA sequencing technologies. Adapted from Gupta and Gupta (2014)

Roche 454, Illumina/Solexa, and ABI-SOLiD technologies are called next-generation sequencing (NGS) or second generation sequencing. In these technologies, all genome/ transcriptome is fragmented, and the resulting fragments are sequenced at the same time – parallel sequencing, still existing the need to amplify them by Polymerase Chain Reaction (PCR) reaction before sequence determination, which is achieved by fluorescence detection. Third generation sequencing (like PacBio and Helicos) is able to define the DNA sequence using single DNA molecules, without need of PCR reactions. IonTorrent

technology is not a single molecule sequencing technique, but neither a second generation hence the determination of sequence does not depend on fluorescence bun on hydrogen ion detection when nucleotides are incorporate into a growing DNA template (Gupta and Gupta, 2014). Some of the technical specifications of these NGS technologies can be consulted in **Table 1.1**.

Other technologies (Gupta and Gupta, 2014) are being developed, using other type of chemistries when sequencing, but the final goal is to make sequencing cheaper and available for everyone. The next generations sequencing may have many applications (Gupta and Gupta, 2014), as whole genome sequencing/re-sequencing, transcriptome sequencing (RNA-seq), sequence of DNA that interacts with proteins, .etc.

Table 1.1: Technical specifications of NGS technologies. Adapted from Gupta and Gupta (2014); Liu et al. (2012); Quail et al. (2012); **www.allseq.com** - accessed on 11 August 2014.

	454	**Illumina (HiSeq 2000)**	**Pac Bio (RS II)**	**Ion Torrent (PMG 318)**	**SoLiD**
Chemestry	Pyrosequencing	Sequencing by synthesis	Real time sequencing	Proton detection	Sequencing by ligation
Machine price	∼$500k	∼$690k	∼$700k	∼$50k	∼$495k
Sequencing cost per million bases	$10	$0.07	$1	$2	$0.13
Read length	up to 1000bp	∼1500bp	3000∼15 000 bp	up to 400b	∼85bp
Observed error rate	∼0.1%	0.26%	12.86%	1.71%	0.06%
Time per run	23 hours	3-10 days	3 hours	4-7 hours	8 days
Total output per run	700 Mb	600 Gb	∼300 Mb	1Gb	120Gb
Advantage	Fast and large read length	High throughput	No PCR nedeed; long reads	Fast	High accuracy
Disadvantage	Hing error introduction in homopolymers	Short read	High error rate	Short reads and low amount of output data	Short reads and low amount of output data

1.2.1 RNA-seq

RNA sequencing (RNA-seq) uses next generation sequencing technologies in order to analyse all ribonucleic acid (RNA) molecules of one or more cells – transcriptome, providing information about their sequence and transcription level, facilitating thus the design of therapeutics and experimental data interpretation (Chu and Corey, 2012), as well as revising previously annotated genes (Nagalakshmi et al., 2010).

The RNA-seq process includes three main steps in order to obtain biological insights from the experiment, ***(1)*** library preparation, ***(2)*** sequencing and ***(3)*** bioinfor-

matic/computational biology analysis of the data, as observed in Figure 1.3.

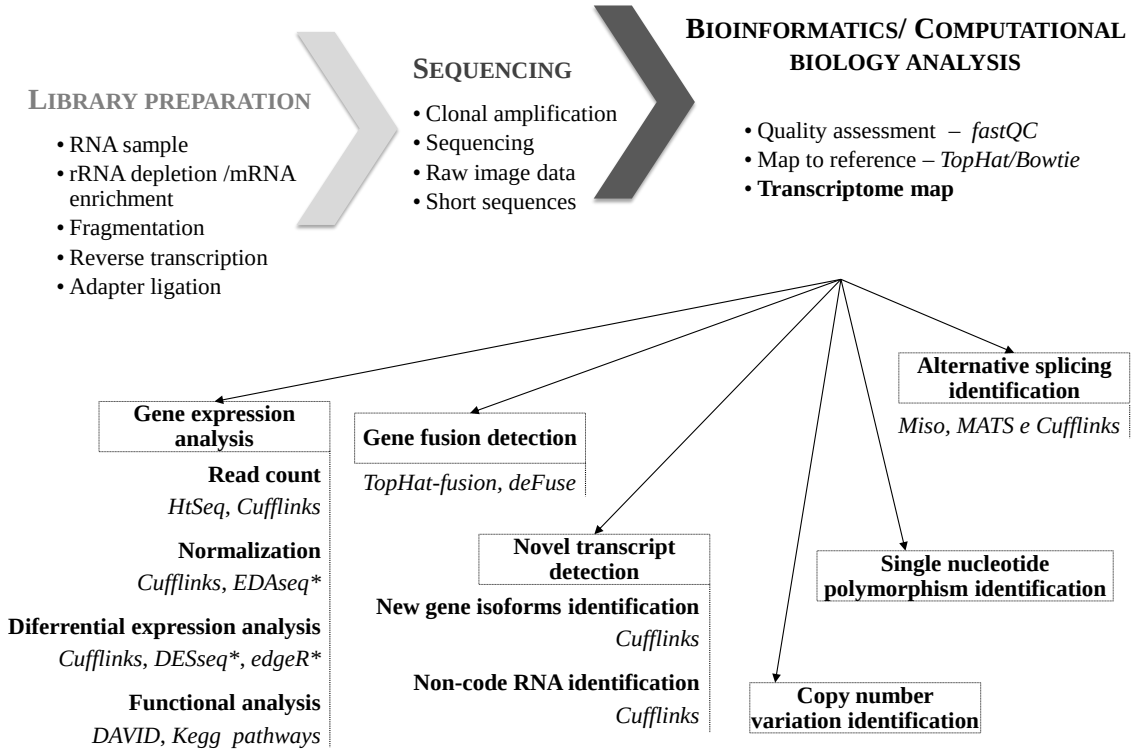


Figure 1.3: RNA-seq process includes three main steps: ***(1)*** library preparation, ***(2)*** sequencing and ***(3)*** bioinformatic/computational biology analysis of the data. Each one of the steps have different sub-steps that need to be followd in order to complete the experiment with success. In the third step are depicted some of the analysis that can be done with this kind of data; as well as some of the open source software available, written in italic and the ones with "*" are R packages that can be installed and do the desired analysis. Adapted from Qian et al. (2014); Atak et al. (2013)

The first step of the process - **library preparation**, involves the conversion of the cellular RNA molecules (total RNA, mRNA or other sub-population of RNA molecules) into molecules that can be sequenced. As ribosomal RNA (rRNA) is one of the most abundant RNA form the cell, these needs to be depleted in order to ensure that transcripts that are rare to be sequenced with a proper depth. Next follows RNA fragmentation into smaller fragments suitable for sequencing, conversion of the RNA fragments into complementary DNA (cDNA), by reverse transcriptase enzyme action, and finally the attachment of adaptors that allow the amplification and sequencing step (Chu and Corey, 2012).

When creating the library, it is possible to retain information of transcript orientation, although more complicated; and during the **sequencing**, the fragments present in the library can be sequenced from one – single-end sequencing, or both ends – paired-end sequencing. The advantage of using paired-end sequencing is that the distance between each paired-end is known, which makes the mapping, against reference annotation, of

reads among repetitive regions more precise (Illumina Inc., 2013). Depending on the used technology, the resulting short reads can cave between 30 and 400 base pairs (bp) of length (Wang et al., 2009; Nagalakshmi et al., 2010).

The usual **analysis work flow** of RNA-seq data (Mutz et al., 2013), if a genome/transcriptome reference is available, consists in:

1. Converting row image data into read sequences;
2. Map read sequences against the reference, in order to reconstruct the transcriptome;
3. Count reads and calculate gene expression and normalizing by region length and library size (RPKMs);
4. Determine differential gene expression using statistical tests.

Several other approaches can be used in order to achieve more specific goals and several tools have been developed in order to permit that (Figure 1.3), most of them being available for free. RNA-seq can offer several advantages (Wang et al., 2009; Nagalakshmi et al., 2010), as it can be used with non-model organism that do not have already genome defined; can reveal the exact location of exon/intron/gene boundaries, gene fusion and sequence variation; identify splice variants and transcription start site; and it is accurate for quantifying gene expression on a genome wide scale.

1.3 Long intergenic non-coding RNA

The non-coding genome regions may function as substrate for DNA-binding proteins (Cheetham et al., 2013) and also as template for the transcription of a vast number of non-coding RNAs (Cheetham et al., 2013; Carninci et al., 2005).

With the development of NGS technologies, a bigger number of transcriptomes have been analyzed, thus acknowledging that a transcriptome is not only constitute by protein coding transcripts but also by a high number of non-coding transcripts, that are being transcript form regions that were believed to be “deserts” or even derived from a protein coding gene. The widespread transcription of non-coding regions may be in the origin of functional molecules (Ecker et al., 2012), making apparent that there is a need to take into account non-coding elements when genome wide association studies are done.

Non-coding RNA (ncRNA) are associated with plenty of functions (Cech and Steitz, 2014) as regulation of gene expression at transcription, RNA processing and translation level, protection of genomes from foreign nucleic acids and can be grouped in two major classes, based on transcript size (Guttman et al., 2011):

- **Small non-coding RNAs** – smaller than 200 nucleotides (nt);

- **Long non-coding RNA** (lncRNA) – bigger than 200 nt and up to 100 kb.

This **200 nt metric** is a convenient cut-off value to exclude small RNAs in purification protocols (Chen and Carmichael, 2010; Derrien et al., 2012; Shi et al., 2013).

According to their location to the nearest protein-coding gene, lncRNAs can be categorized as: **exonic**, **intronic**, **overlapping** or **intergenic** (Guttman et al., 2011; Derrien et al., 2012) and can be found in sense/antisense orientation (Guttman et al., 2011).

Long intergenic non-coding RNAs (lincRNA) , have no overlap with coding genes and can be transcribed from thousands of loci in mammalian genomes (Ulitsky and Bartel, 2013), however very few of the identified ones have been characterized; existing studies that show that these are regulated during development and in response to certain signals, existing also the possibility for being misexpressed in tumors, as reviewed by Tsai et al. (2011).

In the last years the interest in lincRNAs has augmenting, as can be observe in **Figure 1.4**, where is depicted the evolution of the number of results when searching for "lincRNA" since 1976 until nowadays.

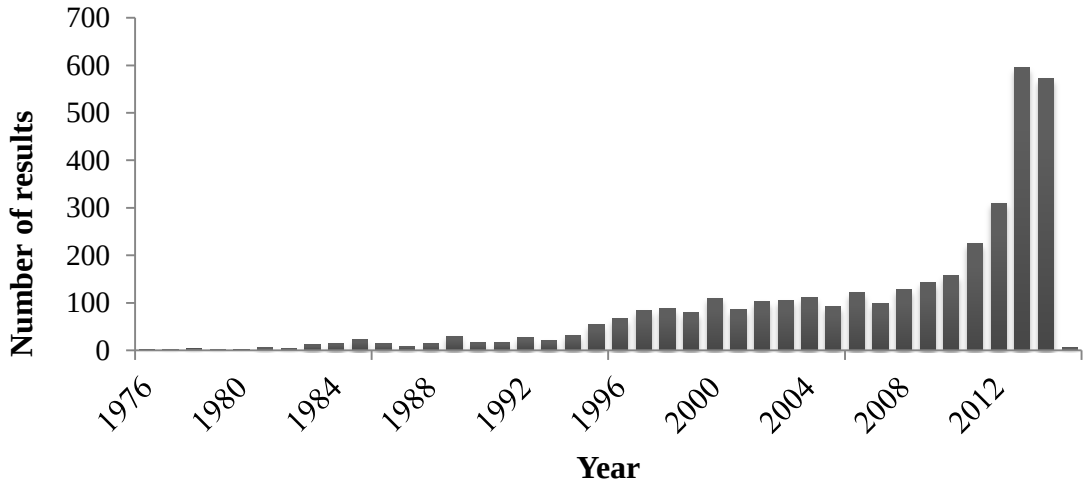


Figure 1.4: Evolution of lincRNA results in **PubMed** since 1976 until nowadays. Data obtained from **www.ncbi.nlm.nih.gov/pubmed** in 21st November 2014.

LincRNA identification

With the methodological advance and with the amount of data produced via NGS technologies, better collections of lincRNA are being produced with accurate genomics positions describing the transcripts; existing nowadays lincRNAs catalogued in several species including human, mouse, zebrafish, fly, nematode and *Arabidopsis*, as reviewed by Ulitsky and Bartel (2013).

Guttman et al. (2009) and Khalil et al. (2009), using chromatin signature of actively transcribed genes - histone H3 trimethylation of lysine 4 in the promoter region and lysine 36 along the transcribed region (H3K4me3 and H3K36m3, respectively) - which characterizes the regions where the polimerase II transcription initiation and elongation occurs, were able to identify about 1600 and 3300 lincRNAs in mouse and human cells, respectively. The lincRNA identified using this approach show strong purifying selection in their genomic loci, exonic sequences and promoter, and most of them evolutionary conservation, therefore presenting functional capability.

Other approaches using a combination of several techniques (RNA-seq, cDNA and chromatin marks) have been being used, with different criteria of certain impositions as transcript size and distance from protein coding gene the consequence being the lack of overlap between different studies lincRNA lists, as hypothesized by Ulitsky and Bartel (2013).

LincRNA genomics

Studies suggest that gene "deserts" near transcription factor genes preferentially hold lincRNA. The main causes for this being the regulation of lincRNA in *cis*, coregulation of lincRNA and transcription factors or the presence of enhancer elements near transcription factors favoring the emergence of lincRNA genes as suggested by Ulitsky and Bartel (2013).

LincRNAs have no sequence or structure characteristic defining them, sharing features with protein coding genes transcripts, hence are frequently transcribed by RNA polymerase II (Guttman et al., 2011) as well as cis/trans-regulatory behaviour (Guttman and Rinn, 2012). In human catalogues lincRNA genes have about 2-3 exons and these are tendencially longer than exons of protein coding genes. They present chromatin modification pattern, transcriptional regulation and splicing sites similar to the ones of protein coding genes; most of the annotated lincRNAs being polyadenylated just like protein coding genes and some of them can even present circular isoform, as reviewd by Ulitsky and Bartel (2013).

LincRNA expression tends to vary between tissues but have a more tissue/cell-specific patterns (Derrien et al., 2012; Cheng et al., 2013; Ulitsky and Bartel, 2013). On average lincRNA expression is about a tenth of the median mRNA level (reviewed by Ulitsky2013), not being known if this is due to less efficient transcription or more efficient degradation. Although the most studied lincRNA (Xist, Malat1 Neat1 and Miat) are found almost exclusively in the nucleus, other lincRNA can be found in the cytoplasm.

Mechanism of action

The biological role of most lincRNA is still unknown, but several mechanisms are suggested based on known cases (Figure 1.5). Ulitsky and Bartel (2013) divides lincRNA

function in three groups:

- The ones that who's nascent RNA is functional, having their action in *cis*, thus the target of the lincRNAs is in the proximity of the lincRNA gene;
- Those who require a processed RNA but still act on *cis*;
- The ones independent of the transcription site – acting in *trans*.

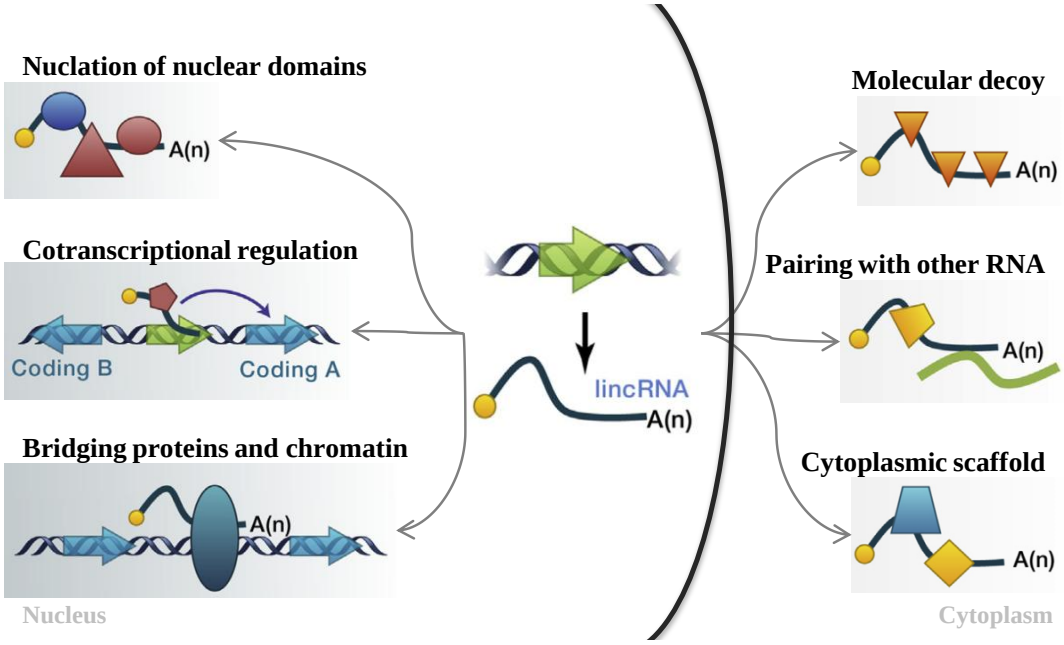


Figure 1.5: Generalized mechanisms of lincRNAs action. Image adapted from Ulitsky and Bartel (2013).

With this technological development, lincRNAs (and all others lncRNA) have received increased attention (Qiu et al., 2013) and are emerging as important regulatory molecules in tumor-suppressor and oncogenic pathways (promoting matrix invasion of cancer cells and tumor growth) , in addition to proteins and microRNAs (Huarte and Rinn, 2010), bringing new potential biomarkers (Cheng et al., 2013) for diagnosis, prognosis metastasis and prediction of therapy response.

1.4 Clear cell renal cell carcinoma

Kidney cancer or **renal cell carcinoma** (**RCC**) is a common group of chemotherapy-resistant diseases, and one of the most lethal type of cancer in the urinary system (Zeng et al., 2014).

Early stages of this type of carcinoma usually does not present any symptoms, but as the tumor progresses symptoms like blood in the urine and pain or lump in the lower

back or abdomen (American Cancer Society, 2014), among others, may appear. The survival rate for patients suffering from metastatic RCC is less than 10% survive five years subsequent to diagnosis (Zeng et al., 2014).

According to the American Cancer Society (American Cancer Society, 2014), surgery continues to be the primary treatment and ablation therapy (uses cold or heat to destroy tumor) can also be performed for patients that are not surgical candidates. This carcinoma exhibits resistance to chemotherapy and radiation but the improvement in acknowledging this carcinoma biology (genetics, affected molecular pathways, *etc.*) allowed to develop targeted therapies that can be used to treat the metastatic disease.

Based on their genetic characteristics, histological features, clinical phenotype and different responses to therapy, RCCs can be subdivided in several subtypes (Rosner et al., 2010; Linehan et al., 2010; Crumley et al., 2013), one of the most common being **clear cell RCC (ccRCC)** accounting for more than 80% of RCC cases (Ljungberg et al., 2010).

One characteristic ccRCC cells, as other cancer cells, is the **metabolization of glucose** (in order to obtain adenosine triphosphate – ATP, used in as energy in different cell processes) mostly **via glycolysis followed by lactate production**, opposed to glycolosis followed by mitochondrial oxidative phosphorylation, in order to produce ATP; process firstly described by Warburg-**"Warburg effect"**. This characteristic derives mostly due to the inactive VHL gene (Pinthus et al., 2011).

1.4.1 ccRCC characteristics

Histologicaly (see **Figure 1.6.B**) ccRCC is characterized by compact nests of tumor cells with clear cytoplasm separated by vascular tissue and **genetically** by deletions of chromosome 3p segments, inactivation of von Hippel–Lindau (VHL) tumor suppression gene gene by mutation and promoter hypermethylation, gain of chromosome 5q and loss of chromosomes 8p, 9p, and 14q (Rini et al., 2009).

The VHL gene mutations were observed only in 52% of the analysed samples, which may indicate that this mutation is not sufficient to explain this carcinoma and that more studies are necessary in order to understand this carcinoma.

An important role of epigenetic regulation has also been suggested for ccRCC (Larkin et al., 2012; The Cancer Genome Atlas, 2013; Rydzanicz et al., 2013). Several mutaitions have been identified in tumor suppressor genes involved in chromatin and histone modifications in ccRCC, such as: SETD2 methyl transferase , SWI/SNF chromatin remodelling complex gene PBRM1, BAP1 histone deubiquitinase and KDM6A and KDM5C demethylases, among others.

(a) (b)

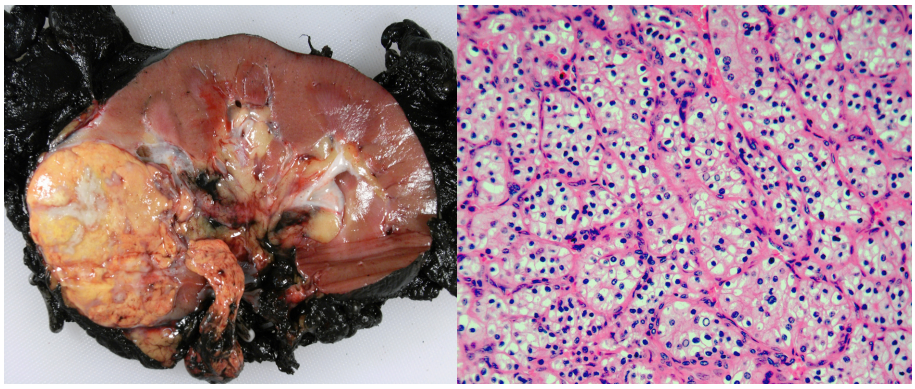


Figure 1.6: **(A)** Typical presentation of a kidney with ccRCC. Yellowish color in the lower left corner due to lipid accumulation. **(B)** Histologic staining with hematoxylin and eosin of ccRCC tumor cells. It is possible to observe nests of epithelial tumor cells with clear cytoplasm and a distinct cell membrane, separated by vascular tissue. Figures adapted from Cohen (2014)

Deregulated microRNAs (small ncRNAs that regulate gene expressions) expression, have also been associated with chemotherapeutic resistance of tumor cells, as reviewed by Rydzanicz and colleagues (Rydzanicz et al., 2013), bringing a new opportunity for discovery of molecular biomarkers, predict outcome and new therapy targets.

Glucose metabolism

VHL is part of the E3-ubiquitin ligase complex that binds to the hypoxia-inducible factor subunit α (HIF-α) under normal oxygen concentration conditions and directs it to **proteosomal degradation**.

Hence **VHL is inactive in ccRCC**, even under normal oxygen concentrations, HIF-α is stable, thus it translocates into the nucleus, dimerizing with (HIF-β) to generate HIF-1. **Increased levels of HIF-1** results in the **upregulation of genes with hypoxia response element** (HRE) site (Pinthus et al., 2011); many of these genes promote adaptation to acute or chronic hypoxia, including the vascular endothelial growth factor (VEGF), which promotes angiogenesis (Audenet et al., 2012).

HIF-1 controls **glucose transporter 1** (Glu1), which encodes for a membrane transporter of glucose, **augmenting the influx of glucose** into cancer cells. Hence the **oncogenic aerobic glycolysis** if favored, glucose is converted into pyruvate that afterwards is reduced to **lactate** by lactate dehydrogenase (LDH) and transported out of the cell by monocarboxylate transporter 4 (MCT4); both LDH and MCT4 being controlled by HIF-1. Lactate production induces several **oncogenes**, creates an **acidic environment** that protects cancer cells from the immune system and has been shown to **stimulate HIF-1-α accumulation**, thus augmenting HIF-1-regulated glycolysis (Pinthus et al., 2011).

HIF-1 also inhibits oxidative phosphorylation by **upregulating pyruvate dehydro-**

genase kinase 1 (PDK-1) (Pinthus et al., 2011) that **inhibits pyruvate's dehydrogenase** catalyzing activity over pyruvate to acetyl-CoA, entry product for oxidative phosphorylation.

The increased oncogenic aerobic glycolysis is also an effect of the **PI3K/Akt signaling pathway** (pathway in renal carcinoma) (Pinthus et al., 2011). This pathway can induce the expression and membrane translocation of **glucose transporter**s, increased **hexokinase expression and activity** (activation of phosphofructokinase 1 that catalyses the first irreversible reaction in glycolysis - phosphorylation of glucose to glucose-6 phosphate). This pathway activity also **increases the activity of mammalian target of rapamycin** (mTOR) (Pinthus et al., 2011), kinase that regulates cell growth, proliferation, motility, survival, protein synthesis, and transcription.

Although clear-cell RCC cells obtain their energy mostly from oncogenic aerobic glycolysis (produces 2 moles ATP per mole of glucose), rather that mitochondrial oxidative phosphorylation (produces 36 moles of ATP per mole of glucose), the rate of oncogenic aerobic glycolysis is sufficient to support the rapid and unrestricted proliferation (Pinthus et al., 2011).

Pentose phosphate pathway

ccRCC cells also present an increased activity of the pentose phosphate pathway (PPP), in which nicotinamide adenine dinucleotide phosphate (NADPH) and ribose-5-phosphate (R5P) are produced (Pinthus et al., 2011).

Via PPP, glucose is converted into ribulose, generating reducing agent **NADPH**, which is critical for **defense against oxidative stress**, **lipid synthesis** and other anabolic reactions, factors that enable ccRCC cells to have **higher resistance to apoptosis**, **oxidative stress** and radiations, when compared with normal cells (Pinthus et al., 2011).

Lipid metabolism

Lipid metabolism is an important process in cancer cells; due to rapid cell proliferation there a need for phospholipids and cholesterol to constitute the new cells cell membrane. In ccRCC, the histological clear cell appearance when stained is due to a high cytoplasmic accumulation of lipids (Rezende et al., 1999) being suggested that this accumulation occurs in order to exhaust the NADPH produced via PPP (Pinthus et al., 2011).

1.4.2 Therapies used in ccRCC

The use of **immunotherapy** using cytokines (IFN-α and IL-2 therapies) for metastatic RCC, implies usage of high doses, associated with toxicities, fact that limited their use (Jonasch et al., 2012).

Three main targets are used as **molecular targets** for ccRCC treatment: PI3K/Akt signaling pathway, mTOR and VEGF, as can be observed in **Figure 1.7**. With an inactivated VHL gene,there is a stabilization and accumulation of HIF transcription factors. This accumulation can also result via **PI3K/Akt signalling pathway**, hence with an increased activity of **mTOR**, p70S6 kinase (p70S6K) is activated, enhancing the translation of certain proteins, including HIF (Audenet et al., 2012). Stable HIF protein, translocates into the nucleus enabaling the transcripton of hipoxia inducible genes as well as **VEGF** and platelet-derived growth factor (PDGF) leading to cell migration, proliferation, and permeability.

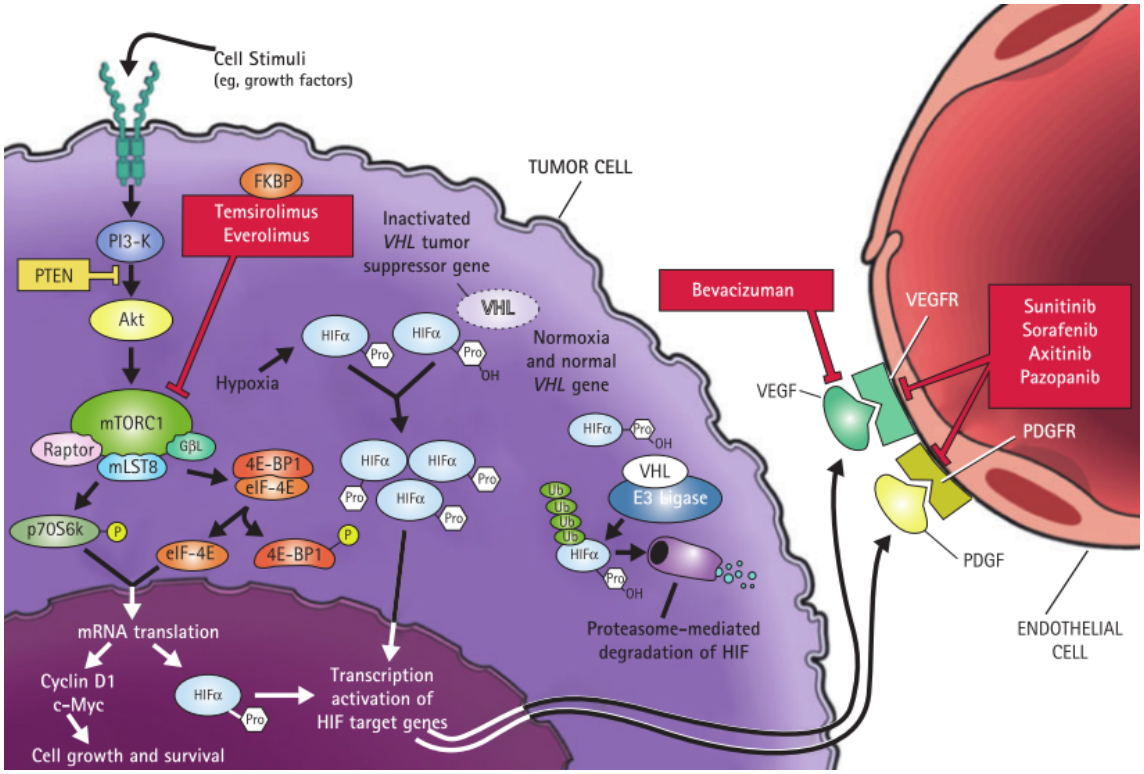


Figure 1.7: Molecular targets of treatments for ccRCC. Inhibition of the kinase activity of mTOR complex 1 (mTORC1) using *Temsirolimus* that binds to FK506-binding protein. *Bevacizumab* is ligand-binding antibody to VEGF (inhibiting its binding and activation of the receptor) and *Sunitinib* and *Sorafenib* are molecule that can inhibit VEGF receptor and PDGF receptor (PDGRF). *PTEN* is phosphatase and tensin homologue that inhibits PI3K/Akt signaling pathway . Pro is a proline residue. Image adapted from Audenet et al. 2012.

As reviewed by Rydzanicz et al. (2013), the use of this this targeted therapy has shown to be more effective and less toxic than immunotherapy but the cost of the drugs are quite elevated, present significant side effects and have relatively low response rates, still being necessary to study in more detail ccRCC's molecular characteristics in order to enable a personalized treatment for each patient, taking into account all the patient and tumor characteristics.

1.5 Objectives

Despite the amount of studies made and the amount of identified mutations it is still not possible to comprehend this subtype of renal carcinoma, hence we decided to explore the lincRNA expression in ccRCC.

This work contemplates to describe how many and which lincRNA are differentially expressed and their correlation with protein coding genes involved in ccRCC cancer.

Specifically, this work aims to:

1. Assemble the ccRCC transcriptome as base for potentially new lincRNA discovery.
2. Analysing differential lincRNA expression to characterize ccRCC.
3. LincRNA correlation with protein coding genes.

In order to achieve this objectives, a human lincRNA catalogue, with lincRNA annotations from several databases has to be constructed in the interest of having a more complete tool/resource resource for assessing lincRNA expression.

Lastly, but not less important, this work focuses in the computational analysis RNA-seq pair-end data of 62 matched normal/tumor ccRCC samples, in order to unravel the lincRNA profile in ccRCC and quantify difference in gene expression when comparing the normal versus the tumor samples.

Chapter 2

Human LincRNA catalog

The method and results presented in this chapter are representative of the collaborative work with Joana Tavares.

2.1 Introduction

To uncover the lincRNA gene profile in ccRCC it is important to use a big compendium of lincRNA annotations. To create this compendium, several databases that contain information about lincRNAs need to converge into a unique annotation that is the most complete and has no redundancies; creating thus a new unified annotation of all lincRNA genes across databases. Hence the ultimate goal is to identify which gene and not which gene isoform may or may not vary, all gene isoforms when existing, were merged and if the same gene exist across several databases, the final version of the gene is obtained by merging all of them into one transcript that covers all gene and exons boundaries across the databases.

2.2 Method

The software used in order to construct the annotation includes **R 3.0.0** and **Bedtools 2.19.0**. The lincRNA list annotation was obtained by combining available data from **Ensemble Gene 74**, **Vega 54**, **Gencode 19**, (Volders et al., 2013), **Lncipedia 2.1** (Xie et al., 2014), UCSC (downloaded on 19th February 2014), **Broad Institute**'s LincRNA catalog (Cabili et al., 2011) - downloaded on 19th February 2014; and **Zhipeng et Adelson** supplementary table S1 information for human data (Qu and Adelson, 2012).

Three crucial steps are necessary to obtain this unified list and these are: (1) *filtering for lincRNA transcripts*, (2) *merging different gene isoforms* and (3) *combining the information present in all different databases* (as observed in **Figure 2.1**). This steps are further described in the following subsections of this Method section.

All downloaded data and the final list are converted into standard format files for exons

and transcripts - **.gtf** and **.bed**. To ensure that there are no mistakes in the files, all files were uploaded into UCSC Genome Browser.

2.2.1 LincRNA obtainment

After downloading the information from the databases, in order to have only lincRNA, it is necessary to filter for:

- Non-coding transcripts;
- Transcripts longer than 200 nucleotides;
- When available, transcripts that are intergenic.

In the case of the information retrieved from UCSC and Lncipedia, the filtering process only allows to obtain lncRNA, being necessary to subtract the Ensembl coding gene information in order to obtain the lincRNAs. All the other lincRNA obtained from the remaining databases will suffer the same process, and to ensure that there is no overlap independently of strand, hence the available RNA-seq data is not strand specific, this subtraction will be independent of strand information.

This subtraction of the coding genes, is achieved by using the Bed tool ***subtractBed***, with the option ***-A*** being used in order to remove a transcript even if it is overlapped with a coding gene by a base pair.

2.2.2 Isoform merge

When existing several isoforms for the same gene and if these are overlapped, after this **second step** there will be only one isoform for a each gene.

This part of the process takes advantage of the Bed tool *mergeBed* with ***-s*** option in order to take into account gene orientation, as well as ***-nms*** to retrieve the names of the merged isoforms, further .bed/.gtf file construction. To ensure that only the isoforms belonging to the same gene were going to be merged, a R script was integrated with *mergedBed* tool. For Zhipeng et Adelson data was used only the *mergedBed* tool with the above specifications, hence all the transcripts were considered as unique genes.

2.2.3 All databases merge

Finally, when merging all lincRNA information from all databases, it is necessary to concatenate all .gtf files that resulted from the previous step, and use *mergeBed* tool, using again ***-s*** and ***-nms*** options as well as ***-d -8*** in order to ensure that the merged exons have an overlap of at least 8 base pairs. After obtaining the merge result it is necessary to correct the names of the exons, in order to ensure the correct construction of the resulting transcript and to achieve that a R script was used.

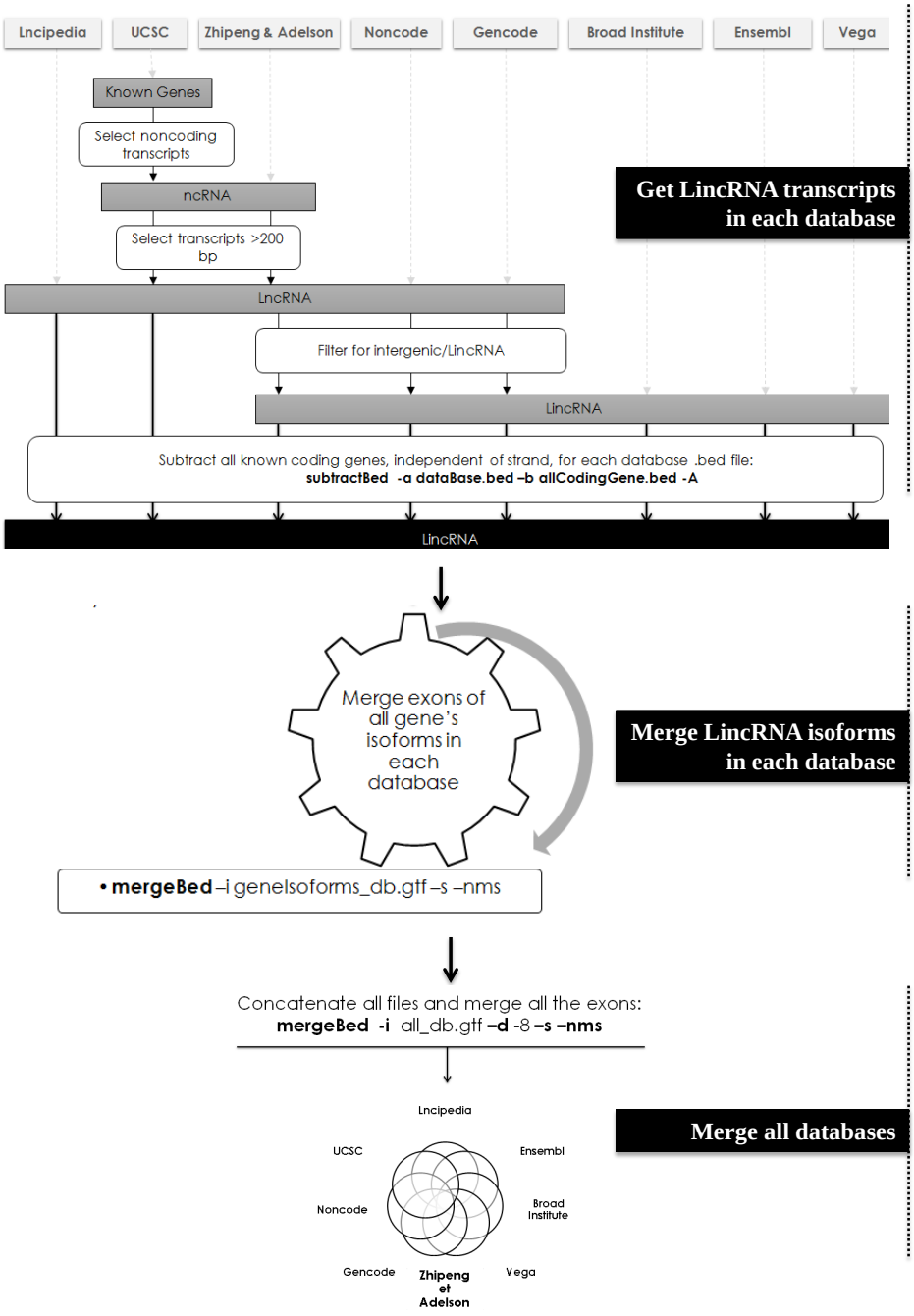


Figure 2.1: Process followed in order to obtain a unified LincRNA annotation accros several databased. This process is composed by three steps that include the obtainment of LincRNA transcripts from all the databases, merging isoforms for all lincRNA genes and lastly the combination of all different databases. **Dotted arrows** lines - original type of information retrieved from the database and the **solid lines** the different processes that the original information suffered.

2.3 Results and discussion

2.3.1 LincRNA obtainment

After downloading the available information several steps were needed in order to obtain only lincRNA information. The criteria, in order to filter what was or not lincRNA, were:

- Transcript must be non-coding - **ncRNA**;
- Transcript bigger than 200 nucleotides -**lncRNA**;
- Be intergenic (no overlap with coding genes) - **lincRNA**.

The data available from Zhipeng et Adelson publication had no gene information, fact that obligate consider all transcripts as individual genes.

To ensure that all lincRNA genes in the annotation were intergenic, independently of their orientation (hence the RNA-seq dataset is not strand specific) **all Ensembl coding genes were subtracted**. In **Figure 2.2** is possible to observe the number of genes lost during subtraction and the associated gene lost percentage, as well as the number of genes after.

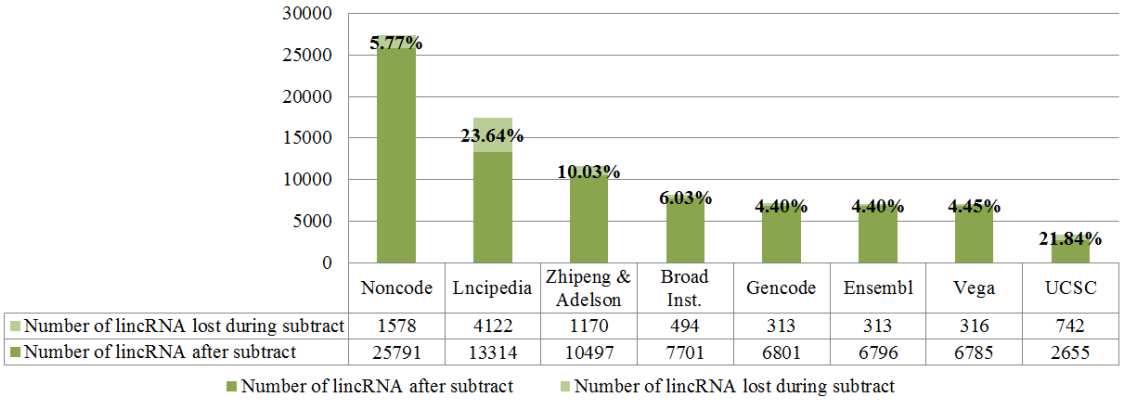


Figure 2.2: Subtracted lnc/lincRNA that are overlapping coding genes, independently of strand.

An interesting and expected tendency during this subtraction is the lost gene percentage of databases that after filtering had lncRNA was similar (around 20%) whereas the lost gene percentage of databases that already contain only lincRNA was equal or below 10%. After this step it was possible to evaluate each database in concordance to their lincRNA content and it was possible to observe that the database with more lincRNA genes/ transcripts is Noncode (as observed in **Figure 2.3.A**) and the database with more isoforms per gene is Lncipedia (as observed in **Figure 2.3.B**). This last figure also allows to see that for all databases more than 70% of genes have only one isoform and most of databases have in average 3 exons per transcript (**Figure 2.3.C**), similar with what is described in literature (Cabili et al., 2011; Derrien et al., 2012; Ulitsky and Bartel, 2013).

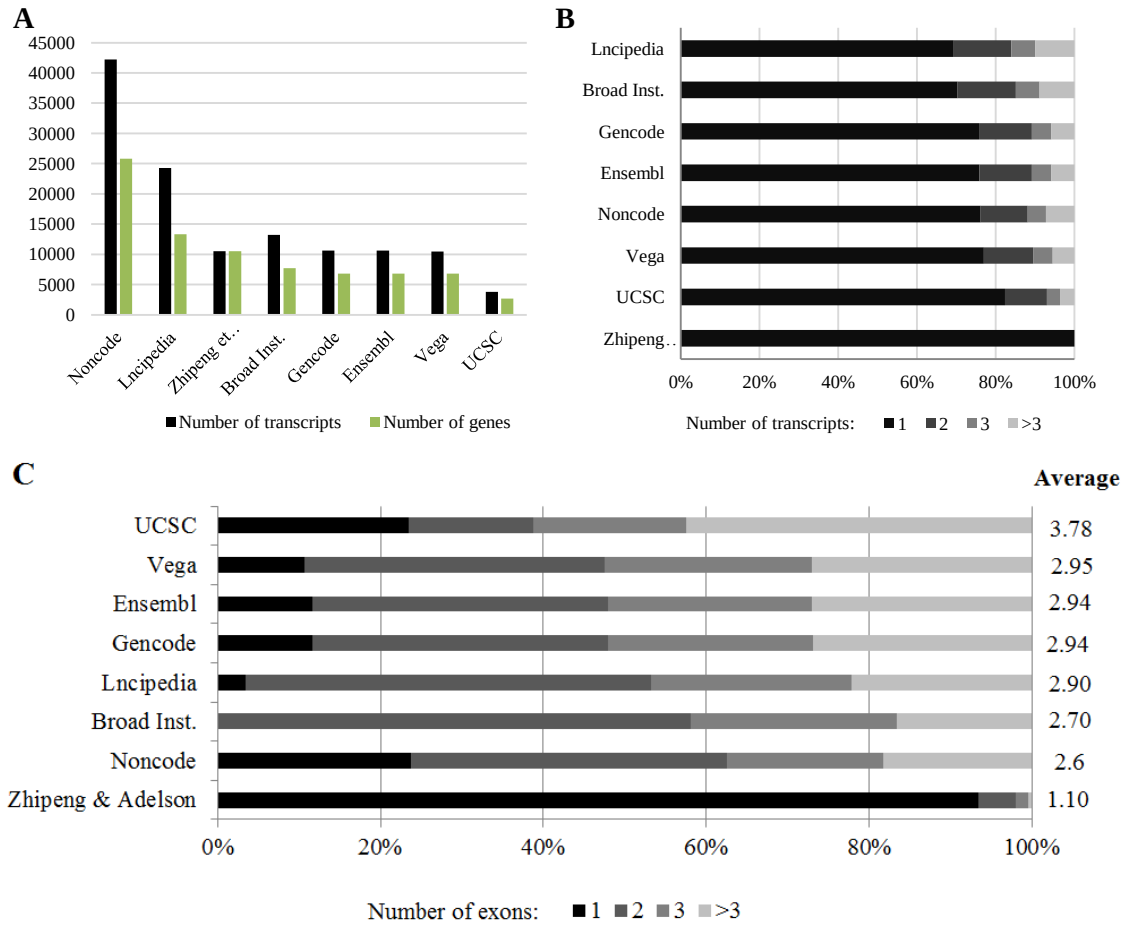


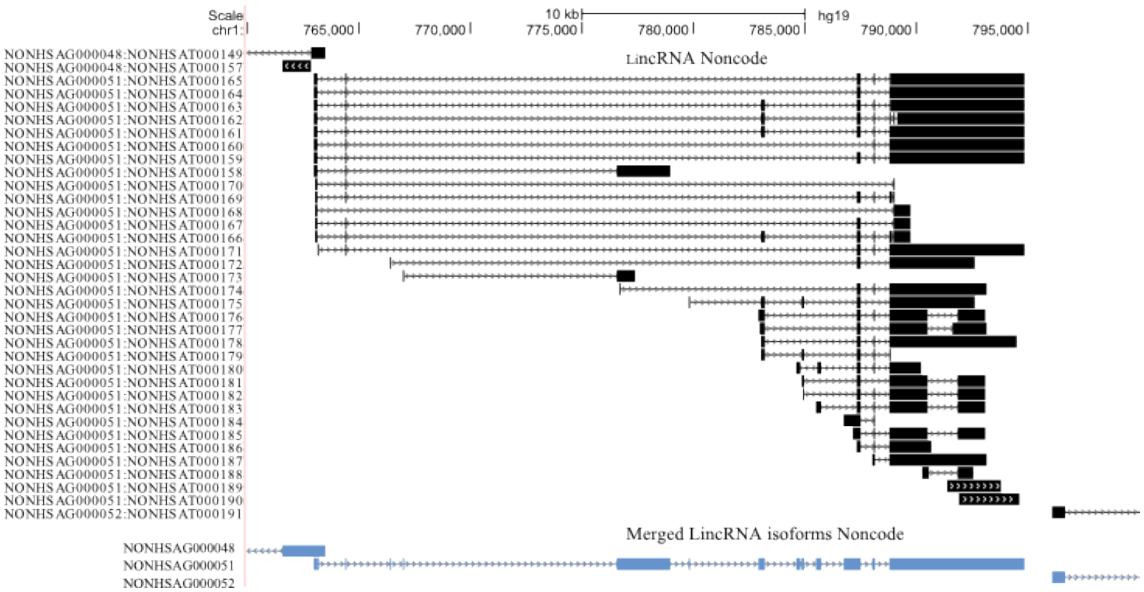
Figure 2.3: LincRNA information available in databases.**(A)** Number of transcripts and genes in each database. The database with more genes transcripts is Noncode whereas the one with least is UCSC. **(B)** Information about how many genes (in percentage) have one or more isoforms. The database with less gene percentage of genes with more than one isoform (excluding Zhipeng et Adelson hence all transcripts are considered as unique genes) is UCSC, whereas the one with most gene percentage of genes with more than one isoform is Lncipedia. **(C)** Number of exons per transcript in each database (in percentage) as well as associated average of exons per transcript.

2.3.2 Isoform merge

Hence the percentage of genes with one isoform is quite high (over 70%) in all databases and the main goal is to identify which lincRNA is up/down regulated in ccRCC, and not which isoform in specific, the isoforms merging in each database was a natural approach to follow.

A good example of what happens during this isoform merge follows (**Figure 2.4.a**). Here we can see a gene from Noncode database that initially (black representation) has about 30 different isoforms and the result after the merge has only one isoform (blue representation).

(a)



(b)

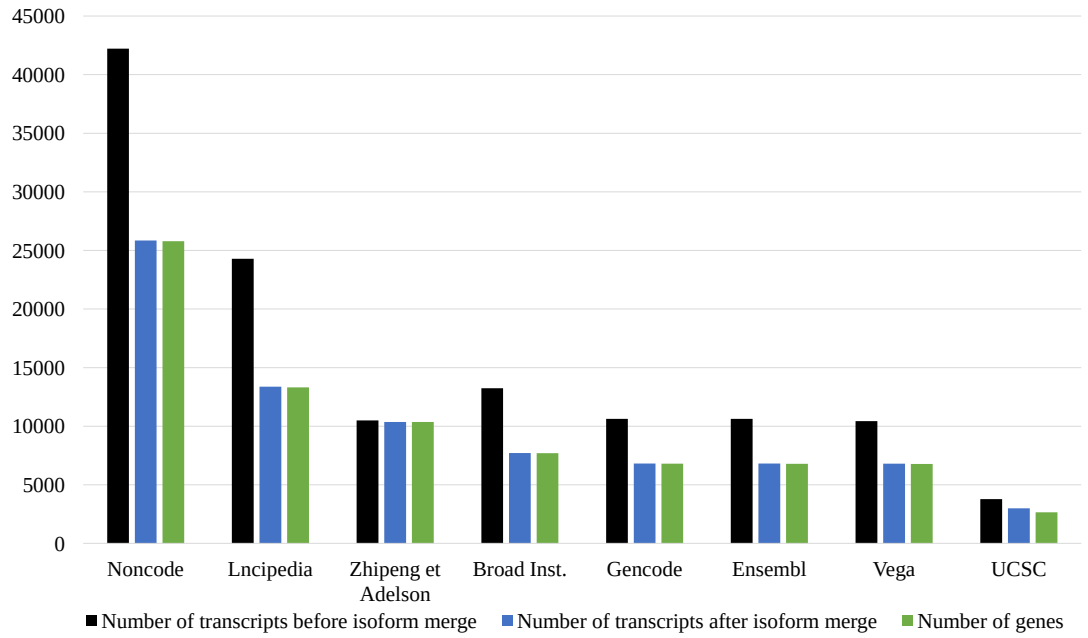


Figure 2.4: ***(a)*** Visual example of what happens during isoform merge. About different isoforms belonging to the same gene - represented in black; after isoform merge are represented by a unique transcript - represented in blue.***(b)*** Number of LincRNA transcripts before and after isoform merging. Number of transcripts after isoform merge is similar to number of genes as expected.

Due to the script construction only isoforms that belong to the same gene and that are overlapped are merged, fact that implies that the gene number remains the same before and after the merge, with the exception of the Zhipeng and Adelson data. Exceptions also exist; due to presence of genes that have isoforms in different chromosomes (mainly in UCSC database) or that are in the same chromosome but at a substantial distance between (isoforms separated by 200kb for example) them and in this cases, where there is

no overlap between the isoforms no merging occurs, the gene continues to present this characteristic even after this **isoform merging process**. In **Figure 2.4.b)** it is possible to observe that the number of transcripts after merge is more similar to the number of genes, and the discrepancies that exists are due to the already approached exceptions.

2.3.3 All databases merge

The merge of all databases makes possible to obtain a unified version of a gene and also to construct the compendium of all lincRNA across several databases. In **Figure 2.5.a** is possible to observe that the same gene has different versions in different databases (represented in blue) and after the final merge there is a unique version of the gene (represented in purple).

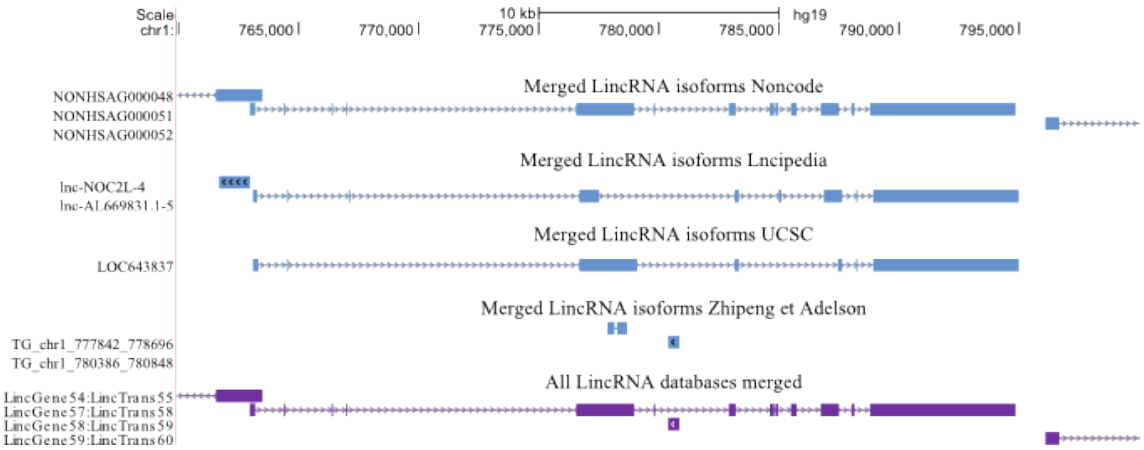
When using this approach, what actually happens in order to construct the final gene, all exons from all different databases are merged as specified and afterwards, in order do consider several transcripts as belonging to the same gene, at least one exon of a transcript from one database needs to overlap an exon from a different database.

The most important fact about this last approach is that we actually gain about 12,000 lincRNA genes when compared to Noncode, database with most genes defined as lincRNAs (**Figure 2.5.b**).

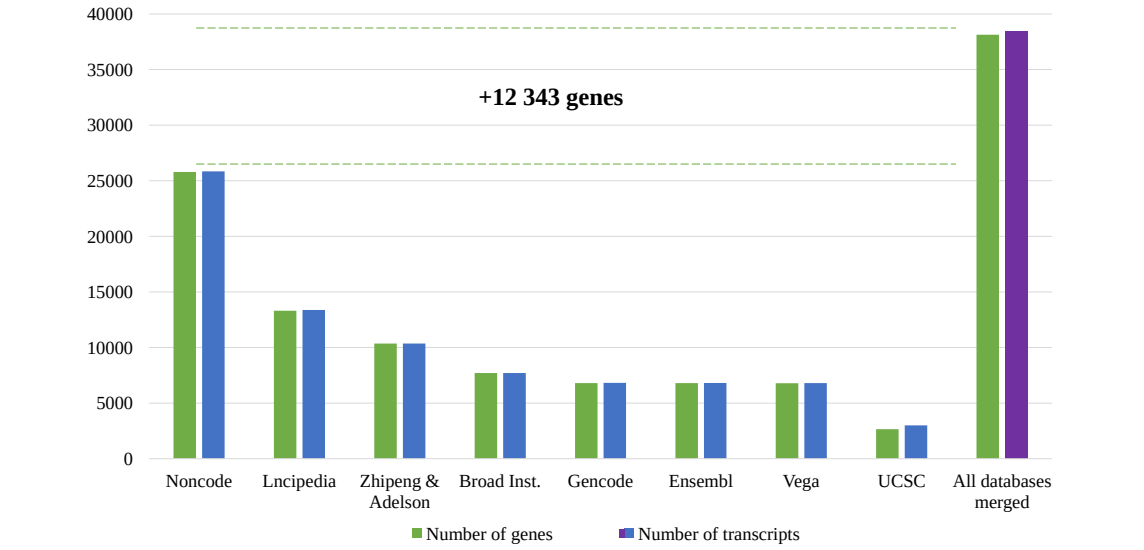
This exon overlap created an unexpected case, where two separated genes from one database (with the same orientation), due to the presence of a gene in another database that overlaps exons of both genes, creates a new transcript that is the combination of these three different genes (as observed in **Figure 2.5.c**). This happens for 1.56% of all transcripts present in this final merge.

For this 1.56% cases and even for all the genes annotated in all databases, in order to assure that the gene structure is correct, transcription data for different tissues and developmental stages, should be integrated with known annotations in order to achieve an optimal definition of each gene.

(a)



(b)



(c)

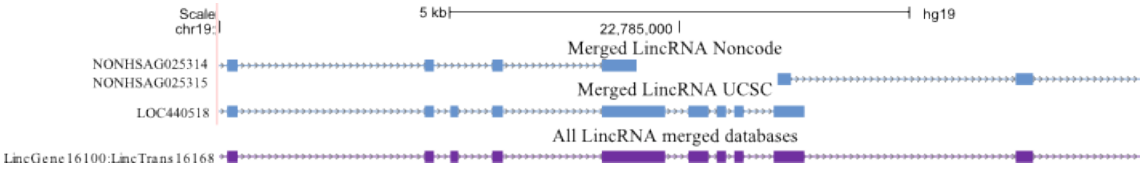


Figure 2.5: **(a)** Visual example of what happens during all databases merge. Results from isoform merging of the same gene in different databases - blue, and the result after all databases merge - purple.**(b)** Number of genes/transcripts across all databases and when combining all together. **(c)** Wrong merge example. This happens for 1.56% of all transcripts present in the final merge.

This final merge also allowed to observe in intersection (**Figure 2.6**) between all databases and in is noticeable that there is a lot of redundancy between them; fact that accentuates the need for a database or an entity that can integrate and regulate, in this case, the lincRNA annotation. Another thing that this figure allows to understand is which databases contributes more with unique genes; here we can see that beyond Noncode (database with more genes), Zhipeng and Adelson data and Lncipedia are the ones that

allowed to increase the gene number in the final merge.

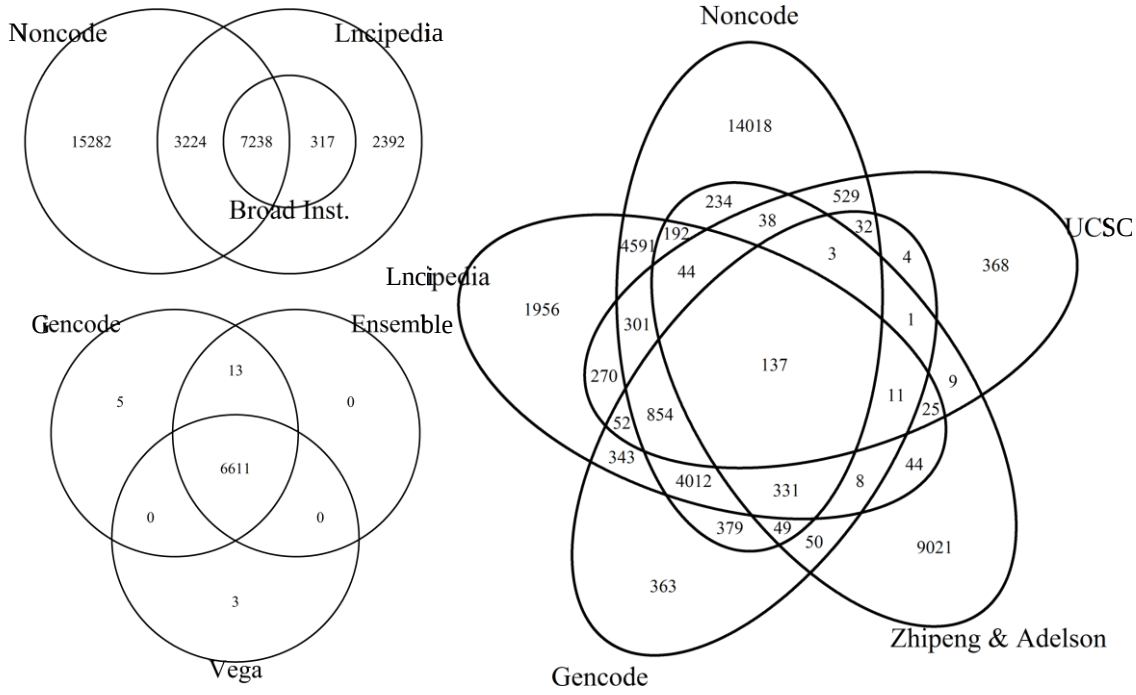


Figure 2.6: Comparison of lincRNA annotations databases

For further analysis along this thesis, this final merged annotation will be used as lincRNAs annotation, being this annotation a unified catalogue of all human lincRNA present across several databases, with a total of 38134 lincRNA genes.

Chapter 3

LincRNA profile in clear cell renal cell carcinoma

3.1 Introduction

3.1.1 Regulatory roles of lincRNA in cancer

Despite the huge efforts of lincRNAs identification, very few of them have been characterized. Several studies show that they are highly regulated during development and in response to signalling and nonetheless can be misexpressed in tumors and leukemias (Table 3.1), as reviewed by (Tsai et al., 2011). The fact that licRNAs can control gene expression and be associated with cancer progression, suggests that their missregulation is not an secondary effect of the cancer; hence the importance in profiling their expression.

Table 3.1: LincRANs associated with cancer. Adapted from Cheetham et al. (2013); Cheng et al. (2013); Huarte and Rinn (2010)

LincRNA name	Cancer type	Molecular mechanism
Over expressed		
ANRIL	Prostate, leukemia	Transcription regulation
HOTAIR	Breast, colon, esophagus, liver, larynx, pancreas	Interaction with PRC2 and LSD1 complexes and targeting to repressed genes
H19	Bladder	Transcription regulation
HULC	Hepatocellular	microRNA decoy
MALAT1	Non-small-cell lung carcinoma	RNA splicing, small RNA production, protein interaction
PCA3	Prostate	Not known
PCGEM-1	Prostate	RNA–protein binding transcription
PVT1	Medeolablastoma multiple myeloma	cMYC-PVT1 fusion protein
Down regulated		
GAS5	Breast, prostate	Decoy of glucorticoid receptor
MEG3	Prostate and others	Not known
PTENP1	Prostate	microRNA decoy
XIST	Breast, ovarian and cervical cancer	Chromatin remodelling

LincRNAs can function through several types of mechanisms, in *cis* or in *trans*, some

of them implying a gene expression control by **recruiting histone modification enzymes** - for example overexpression of HOTAIR, that targets polycomb repressive complex 2, altering histone H3 lysine 27 methylation and gene expression patterns, resulting in an augmented cancer invasiveness; or by **regulating alternative splicing** - MALAT1, as reviewed by Tsai et al. (2011). In Figure 3.1 is possible to observe some of the lincRNA associated mechanisms and some of the associate lincRNAs that use those mechanisms.

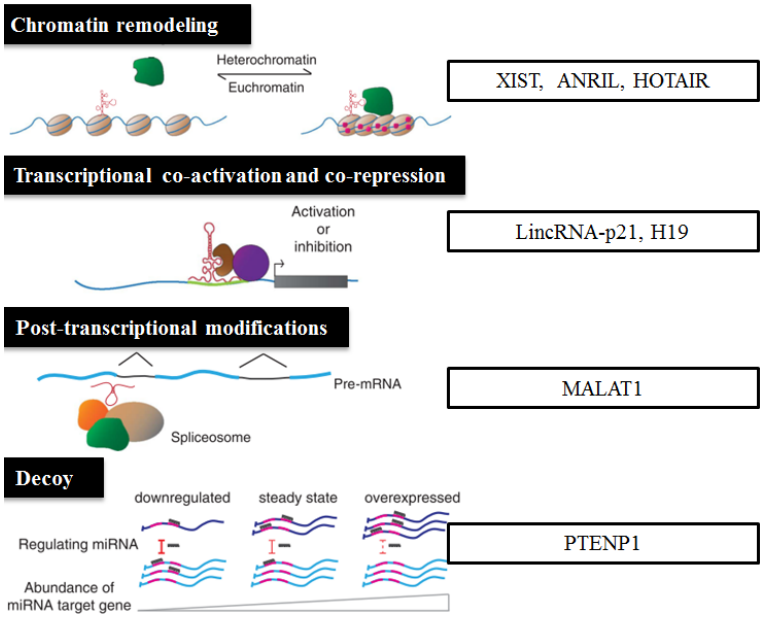


Figure 3.1: LincRNAs associated mechanisms of action and lincRNAs associated with cancer that use that kind of mechanism. Image adapted from Cheetham et al. (2013)

3.1.2 Diagnostic and potential therapeutics of lincRNAs in cancer

LincRNA are emerging as good potential new biomarkers to characterize cancer recurrence, progression and response to therapy (Cheng et al., 2013), hence using only one molecule to infer this type data on a disease is much easier than profiling multiple of other RNA molecules (Tsai et al., 2011). One good example is HOTAIR lincRNA, that can serve as a good independent predictor of eventual metastasis and death (Tsai et al., 2011).

3.2 Methods

3.2.1 Data set

The RNA-seq dataset used consists of 62 matched normal/tumor tissue pairs from ccRCC patients retrieved from The Cancer Genome Atlas (TCGA) Data Portal (//tcga-data.nci.nih.gov). The list with the patient/sample ID for matched normal/tumor samples is available in the appendix **Table A.1.1**. In Table 3.2 it is possible to observe some of the key clinical characteristics of the patients.

Table 3.2: Characteristics of the 62 ccRCC patients whose samples were retrieved from TCGA. The neoplasm histologic grade is relative to cell reversion of differentiation, where G2 is moderately differentiated, G3 is poorly differentiated and G4 is undifferentiated.

Characteristics		# of patients
Gender	Male	44
	Female	18
Age at initial diagnosis	<50	8
	50-60	16
	60-70	22
	>=70	16
Vital status	Alive	40
	Deceased	22
Stage	I	20
	II	9
	III	14
	IV	19
Neoplasm histologic grade	G2	25
	G3	25
	G4	12
Tissue source site	A3 : International Genomics Consortium	2
	B0 : University of Pittsburgh	13
	B2 : Christiana Healthcare	1
	B8 : UNC	5
	CJ : MD Anderson Cancer Center	7
	CW : Mayo Clinic - Rochester	9
	CZ : Harvard	25

The original 124 *.bam* files, used to store pair-end sequence data for all samples, both aligned as well as unaligned reads, were realigned using **tophat** (Trapnell et al., 2012) tool. To realign the reads for all the samples, firstly, each original *.bam* file was converted into two *.fastq* file (format that stores biological sequence of each sequenced read and its

associated per base associated quality score (Cock et al., 2010)) – one per each pair-end, using **bam2fastq** tool (gsl.hudsonalpha.org/information/software/bam2fastq). These two files are consequentially used in the **tophat alignment**, using as reference genome the Ensemble human genome (GRCh37) and the expected (mean) inner distance between mate pairs defined as 150bp. A command example for the realignamet follows:

tophat –r 150 EnsemblIndexGenome fastqFile1 fastqFile2

3.2.2 Transcriptome composition

In order to asses the transcriptome composition of ccRCC samples, **(1)** is necessary create a transcriptome assembly for each alignment obtained from *tophat* (accepted.hists.bam file) using LincRNA annotation combined with Ensemble gene annotation; **(2)** merge all 124 assemblies and finally **(3)** compare the merge result with all Ensembl genes and the LincRNA annotation.

The following commands of **Cufflinks version 2.2.0** were used in order to asses the transcriptome composition:

Table 3.3: Commands used for transcriptome composition assessment

0. Combine annotation between Ensemble genes and LincRNA annotation	
cuffcompare EnsembleGenes.gtf LincAnnotation.gtf	Use cuffcom.combined.gtf as LincAllEnsemblCombined.gtf Use resulting .gtf as annotation for transcriptome assembly
1. Assemble transcriptome for each .gtf file	
cufflinks -r LincAllEnsemblCombined .gtf **-s** genome.fa accepted.hits.bam **-r** Annotation file **-s** Genome fasta files	
2. Merge all transcriptomes	
cuffmerge allGtfPathResultedFromCufflinks.txt	Resulting merged.gtf is the file with all ccRCC sample gene composition
3. Compare known annotation of human genome with transcriptome composition	
cuffcompare -r LincAllEnsemblCombined.gtf merged.gtf **-r** Annotation	Use cuffcomp.combined.gtf exons with class "=" in order to asses transcriptome composition for known genes

New lincRNA discovery

In order to find new lincRNA all exons/transcripts defined as "unknown, intergenic" (class "u") in the file resulting from the above **cuffcompare command** are considered as possible new lincRNA.

In order to asses if these are long non-coding RNA, Noncode functionality iLncRNA (Xie et al., 2014) was used. Briefly, iLincRNA (1) removes transcripts smaller than 200 nt;

the remaining transcripts are filtered by comparison to (2) Ensembl pesudogens and by the (3) coding potential calculation. The remaining transcripts are considered as potentialy new lincRNAs.

The nomenclature used to name the gene/transcript id is the one resulting from **cuffflinks**, "XLOC" plus the "PtNL-"(Potentially New LincRNA) prefix in order to distinguish from possible genes/transcrips from Broad Institute database that have the same "XLOC" nomenclature.

3.2.3 Gene expression

In order to asses gene expression analysis, an annotation file (.gtf file) with Ensembl protein-coding genes, LincRNA annotation and the resulting potentially new lincRNA from Noncode was created, using **cuffcompare** command from Cufflinks version 2.2.0 .

Gene expression quantification (fragments per kilobase of exon per million fragments -FPKM) for each gene present in the combined annotation and the gene differential expression, were obtained with *cuffdiff* test using all 62 normal samples against all 62 tumor samples.

FPKM analysis

All FPKM analysis was conducted in R Studio with R version 3.1.0.

From resulting file "genes.READ_GROUP_TRACKING", were obtained the FPKM values for all genes present in the combined *.gtf* file containing all Ensemble protein-coding genes, licnRNAs and potentially new lincRNA, accros all 124 samples.

Using R, all genes with variation equal to zero were removed from the analysis and, before log2 transformation of the data, all zero FPKM values were substituted with 6.8E-8 in order not to remain with minus infinite values.

Uncovering the FPKM distribution for protein-coding, lincRNA and potentially new lincRNA genes was the first step in understanding how this different genes behaving in the ccRCC transcriptome.

In order to evaluate if it is possible to separate the samples into groups based on gene expression (FPKM values), an unbiased hierarchical clustering of the first 2000 genes, with more variation across all samples, using the Ward's method and principal component analysis (PCA) with all genes expression along all samples were assessed.

To represent the expression of the 2000 genes that most vary, was used a heatmap with green, black and red colors representation; where green is low expressed and red is highly expressed. In order to determine when one color starts and ends color brakes were used that take into account the minimum FPKM value, 25 percentile 75 percentile and the maximum FPKM value.

Differentially expressed genes

From resulting file "gene_exp.DIFF", were obtained the genes that are differentially expressed. The obtaind results were filtered for false discovery rate (FDR) smaller than 0.05 and the absolute value of the log2 fold change grater than 0.58(represents a fold change of 1.5).

3.2.4 Enrichment analysis

In order to proceed the enrichment analysis analysis, differentially expressed (up or down regulated, or a combined list of the previous ones) genes name ("OFFICIAL_GENE_SYMBOL") was used as gene list for DAVID -(http://david.abcc.ncifcrf.gov) - Functional Annotation Tool (DAVID Bioinformatics Resources 6.7, NIAID/NIH), (Huang et al., 2009b, a) input and further, "*Homo sapiens*" was pinpointed as background.

Pathway enrichment analysis was assessed using DAVID Functional annotation tool, only with KEGG_PATHWAY option activated.

Gene ontology (GO) term enrichment analysis was assessed using DAVID functional annotation tool, only with GO_BP_FAT option activated. The visualization of the enriched GO terms obtained from DAVID was done using on-line resource REVIGO - http://revigo.irb.hr/ (Supek et al., 2011) in order to summarize the long list of enriched GO terms, with *Homo sapiens* database for GO terms. In order to limit the number of enriched GO terms, only GO terms with FDR smaller than 0.05 were used in the REVIGO analysis .

Differentially expressed genes name ("OFFICIAL_GENE_SYMBOL") was used as gene list for DAVID input and further, "Homo sapiens" was pinpointed as background in order to proceed the analysis.

3.2.5 Weighted gene correlation network analysis

A weighted gene co-expression network analysis is a method used in systems biology in order to describe correlation patterns between genes across several samples/experiments. Weighted gene correlation network analysis (WGCNA) is available through the R software package WGCNA (Langfelder and Horvath, 2008) that can be used to find clusters of highly correlated genes and relate them to external traits, among other capabilities. Hence it takes into account the relationship among transcripts expression in a global way it can devolve a far richer set of information than just a list of differentially expressed genes. This approach has been successfully used in order to identify candidate biomarkers or therapeutic targets.

A typical WGCNA analysis includes:

- Construction of a gene co-expression network with thousand of genes, that takes into account the interactions patterns between them, being the correlation used as a

measure of co-expression. The resulting matrix that holds the pair-wise correlation between all genes is also called as *adjacency matrix*;

- Define clusters of highly correlated genes - *modules*, using hierarchical clustering and Dynamic Tree Cut as tools;

- Relate modules to external information as clinical data, SNPs, proteomics, *etc.* in order to find biologically interesting modules;

- Studying the relationship between modules;

- Find the key diver genes in the modules with interest. in order to identify candidate biomarkers.

For the weighted gene co-expression network analysis, the R package WGCNA version 1.41-1 was used, taking advantage of the existing tutorials present on CoexpressionNetwork/Rpackages/WGCNA/Tutorials page.

Data preparation for network construction

Hence the data is the result from RNA-seq, the author advises a log2(FPKM+1) transformation in order to not remain with minus infinite values. All 124 samples were used in the construction of the same network. Firstly, sample outliers were removed separately from normal and tumor samples, using WGCNA *flashClust* function to hierarchically clustering the samples based on their expression using the "*average*" method and the *cutreeStatic* function to cut the cluster dendogram with minimum number of object on a branch equal to 10, being the cutoff values 160 and 180 respectively. This permit to remove two normal (TCGA-CW-5591-11 and TCGA-B8-5552-11) and three tumor samples (TCGA-CJ-5681-01, TCGA-CZ-5989-01 and TCGA-B8-4619-01), remaining with total of 119 samples.

These 119 samples were further used in order to remove gene with too many 0 FPKM values along the samples or no variance, remaining with a total of 50852 genes, using for that the *goodSamplesGenes* function.

In order to facilitate the computational analysis, from the 50852 remaining genes, only the first 45000 with more variation were used.

Network construction

Before network construction is necessary to chose a "soft power" in order to emphasize high correlation values at the expense of low values, using the WGCNA *pickSoftThreshold* function; the best power for the present data being 2.

The network was constructed using WGCNA **blockwiseModules** function with Spearman correlation, to take advantage of its rank classification, since the lincRNA gene expression is significantly lower than the protein-coding genes one, in order to find some correlation between these ones expression. The network type is ”signed” which means that the adjacency is calculated as follows:

$$adjaceny = (0.5 \times (1 + correlation))^{softpower}$$

Other parameters used are involved with the modules size definition and this were defined as followed: minimum block size equal to 45000, minimum module size equal to 10, reassign threshold equal to 0, merge cut height equal to 0.25 and deep split equal to 0. For more information on these specifications, method details should be consulted.

3.3 Results and discussion

3.3.1 Transcriptome composition

After merging all assembled transcriptomes for the 62 matched normal/tumor samples and comparing them to the Ensembl gene annotation combined with the LincRNAs annotation it was possible to observe the proportion of protein-coding and lincRNA as well to find potentially new lincRNA. In **Figure 3.2** it is possible to observe the proprotion of each one of them in the ccRCC transcriptome, with a clear and unexpected prevalence of lincRNA (about 50% of the transcriptome is composed by lincRNA's).

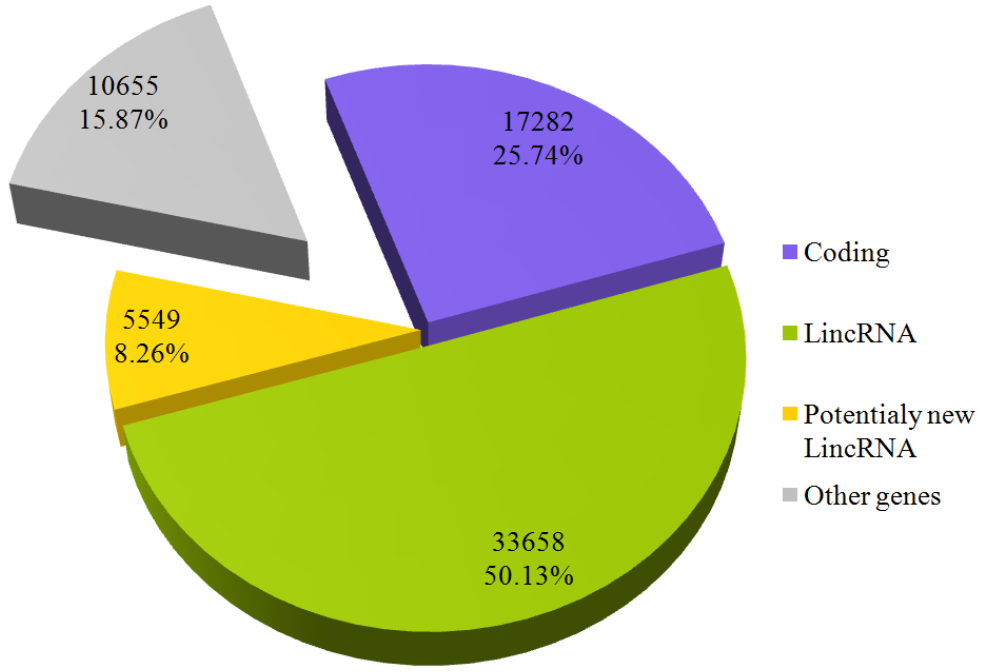


Figure 3.2: Transcriptome composition.

When we take a clear look at what happens in terms of numbers(Table 3.4) becomes clear that although in terms of number of genes, lincRNAs have a higher proportion when we look at the transcripts numbers, protein-coding genes clearly dominate this part, in part, consequence of the form the lincRNA annotation was constructed.

	Protein-coding genes	LincRNA	Potentialy new lincRNA
# genes	17282	33658	5549
# transcripts	56846	33816	5912
# exons	399228	79608	7337

Table 3.4: Trasncriptome composition in terms of the number of protein-coding, lincRNA and potentially new lincRNA genes/transcripts/exons.

3.3.2 Gene expression

Gene expression is evaluated taking into account FPKM values. After their normalization it is possible to observe (Figure 3.3) and better compare the different gene expressions between protein-coding, lincRNAs and potentially new lincRNA genes. When loking at the FPKM values distribution and density, it becomes clear that, although the bigger lincRNAs gene content in the transcriptome, their expression lever is clearly inferior, as reviewed by Ulitsky and Bartel (2013).

The FPKM values distribution and density of potentially new LincRNA is slightly different - higher 75 percentile, lower density of FPKM values close to 0 (log2(FPKM) close to -23) and higher density of FPKM values close to those similar to protein-coding gene expression - when compared to already known lincRNA. This fact is due to the method used in order to discover them - the construction of ccRCC transcriptome and their expression along the samples in order to be defined as "*potentially new/unknown*" by the **Cufflinks** tool. Other hypothesis may be that these are be more specific and related with kidney tissue/ccRCC, as hypothesized by Derrien et al. (2012), when saying that lincRNA expression pattern is more tissue/cell-specific.

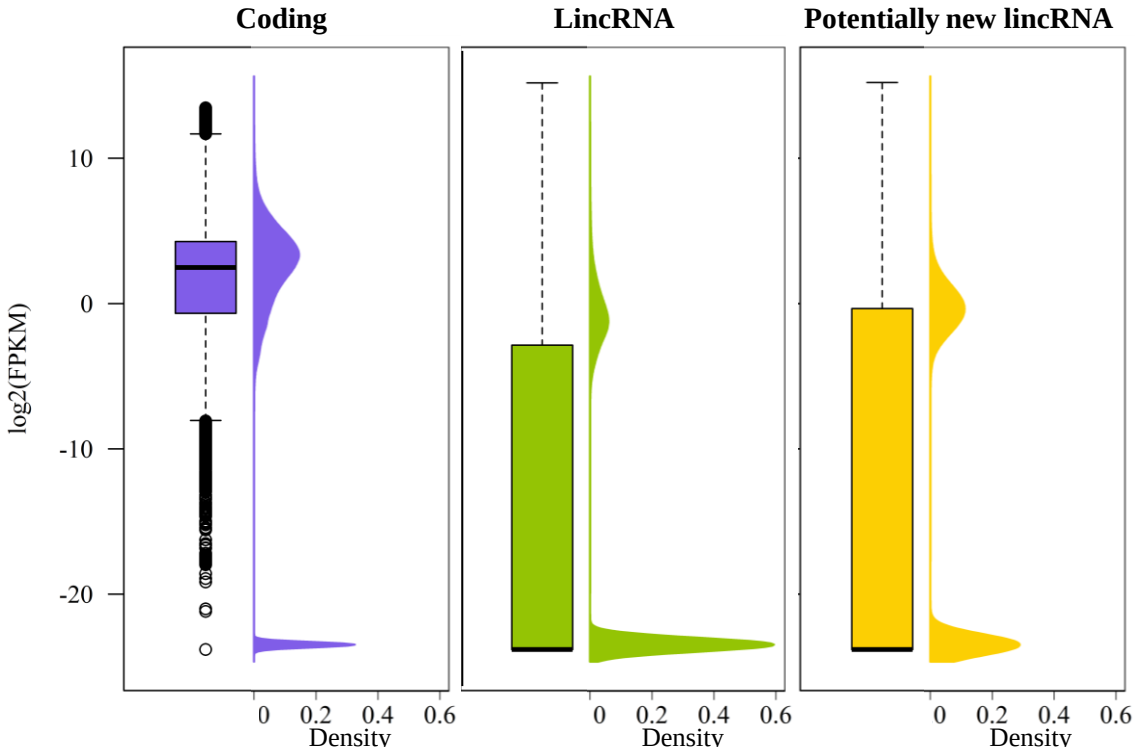


Figure 3.3: FPKM distribution along all samples and all genes with variation bigger than zero, taking into account if genes are protein-coding, lincRNA or potentially new lincRNA.

An unbiased clustering of the FPKM values enables to separate normal samples from tumor samples, as observed in Figures 3.4 *(a)*,*(b)* and *(c)*.

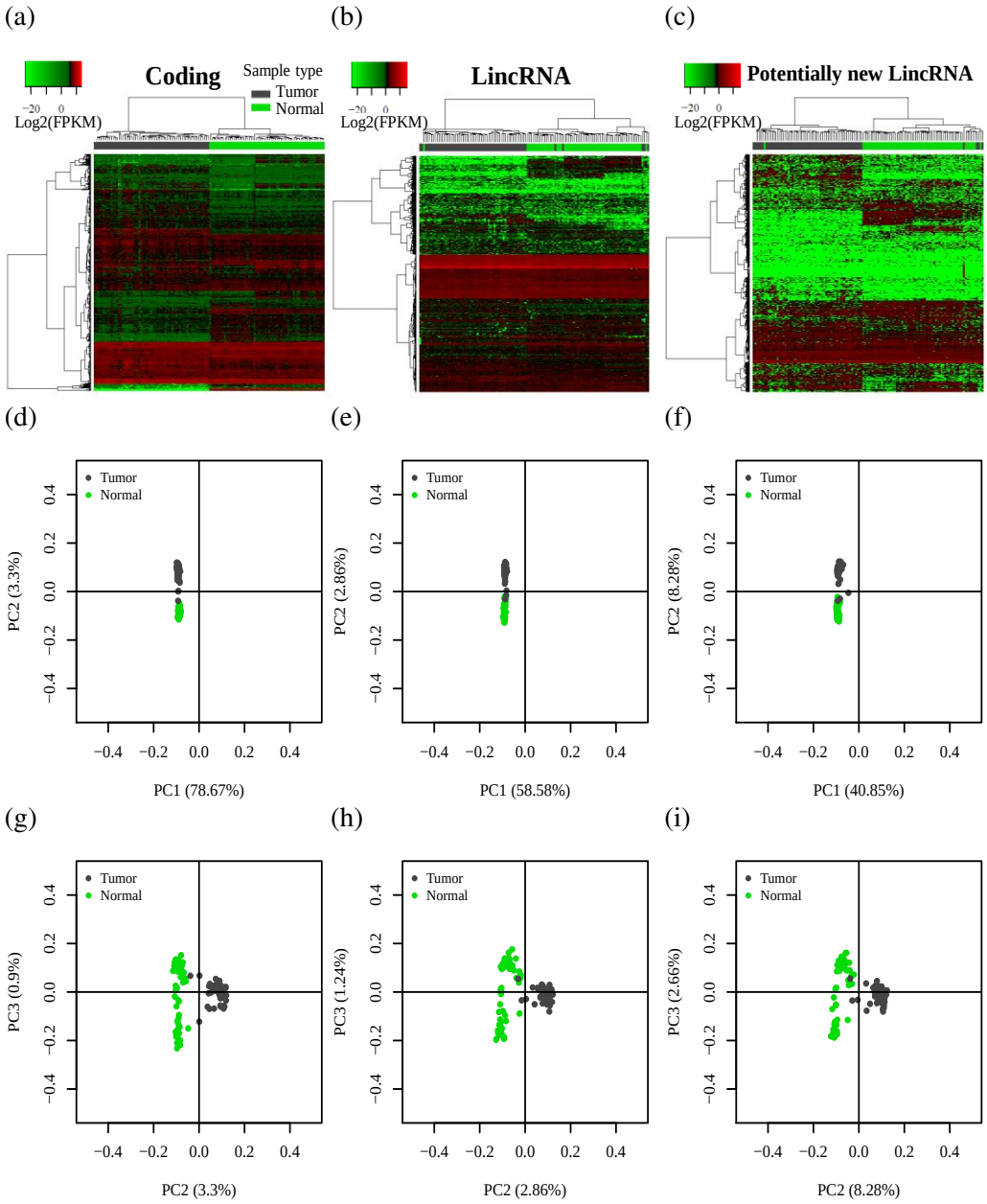


Figure 3.4: FPKM unbiased clustering analysis. Each column is representative of one of the following groups of analysis: protein-coding, lincRNAs and potentially new lincRNAs genes. Each row shows the analysis results. Figures *(a)*, *(b)* and *(c)* show the heatmap of the first 2000 genes with more expression variation in each group. The subsequent 2 rows are respective to the PCA analysis using all genes which expression present a variation above 0. *(d)*, *(e)* and *(f)* show the separation of the samples using first against the second PCA component; *(g)*, *(h)* and *(i)* the second against the third component.

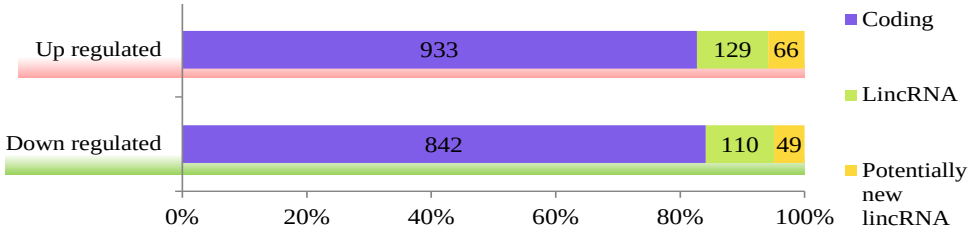
Here we have a clear separation of normal and tumor sample using either one of protein-coding, lincRNA and potentially new lincRNA groups using just the 2000 genes with most variation across all samples. Clear expression patterns can also be observed. When using all genes with variation superior to zero in the principal component analysis

(Figure 3.4 *(d)*,*(e)*,*(f)*,*(g)*,*(h)*,*(i)*) , the first three components also allow a clear separation of this samples; the variation explained by each one of the components is shown between brackets near each one of the components. Although seams to be a clear separation of two groups of normal samples, when separating the samples using principal component 2 against principal component 3, no clear phenotypic characteristic distinguish this samples/patients.

Differentially expressed genes

The differential expression *cuffdiff* test allowed to uncover 2129 genes differentialy expressed, 1128 up and 1001 down regulated in tumor samples when compared with normal samples expression (Figure 3.5 *(a)*). When looking at the proportion of different genes groups it becomes clear that most of them (more than 80%) are protein-coding genes, but the combined proportion of lincRNA and is still significant (less then 20%) and may implicate a role of lincRNAs in ccRCC.

(a)



(b)

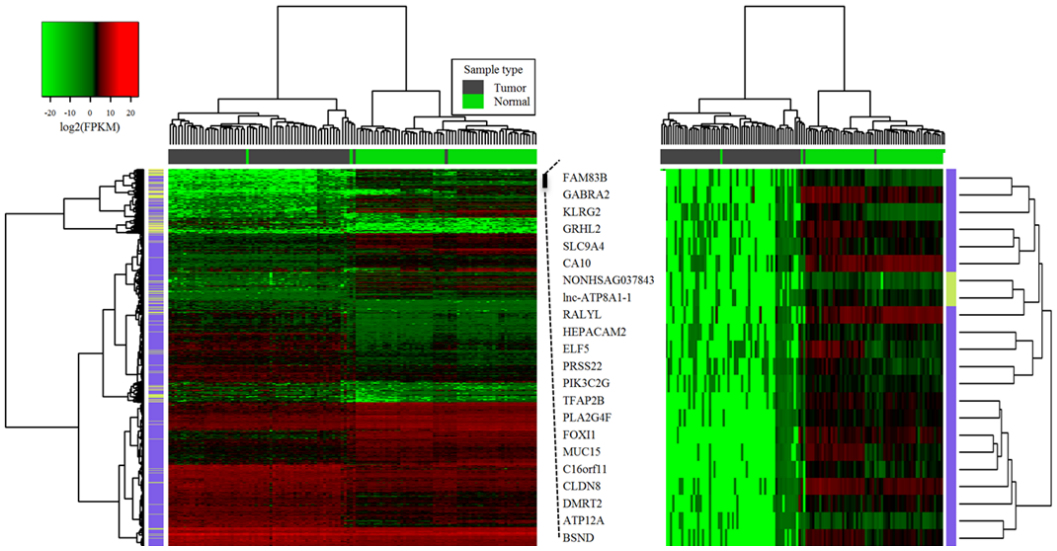


Figure 3.5: *(a)* Proportion of protein-coding, lincRNA and potentially new lincRNA and associated number of up and down regulated genes. *(b)* Heatmap of differentially expressed genes and close up of heatmap in order to show protein-coding gene and lincRNA correlation. Row colors in heatmap represent protein-coding genes, lincRNA and potentially new lincRNA in blue, green and yellow respectively.

Using this differentially expressed genes is also possible to separate most of normal from the tumor samples 3.5 *(b)*), being also possible to observe that there is a correlation between lincRNA expression and protein-coding genes (close-up heatmap).

One lincRNA already associated with cancer is differentially expressed, up regulated - PVT1. The expression of other lincRNAs associated with cancer (per example HOTAIR, PCGEM1, MALAT1, ANRIL, XIST,HULC, PCA3 and MEG3) were not statistically assessed by Cufflinks tool due to low data. This may implicate that a different tool, that is not as sensible to low expression FPKM values, in order to not test their differential expression, should be used in order to evaluate lincRNAs differential expression.

When analysing the differentially expressed protein-coding genes, in terms of their biological process/pathways enrichment, the important roles of the regulation of the response to wounding, leukocite activation, cell adhesion, pyruvate metabolism, glucogenesis (Figure 3.6 *(a)* ,*(b)*) that are highly associated with ccRCC, appear clearly.

(a) (b)

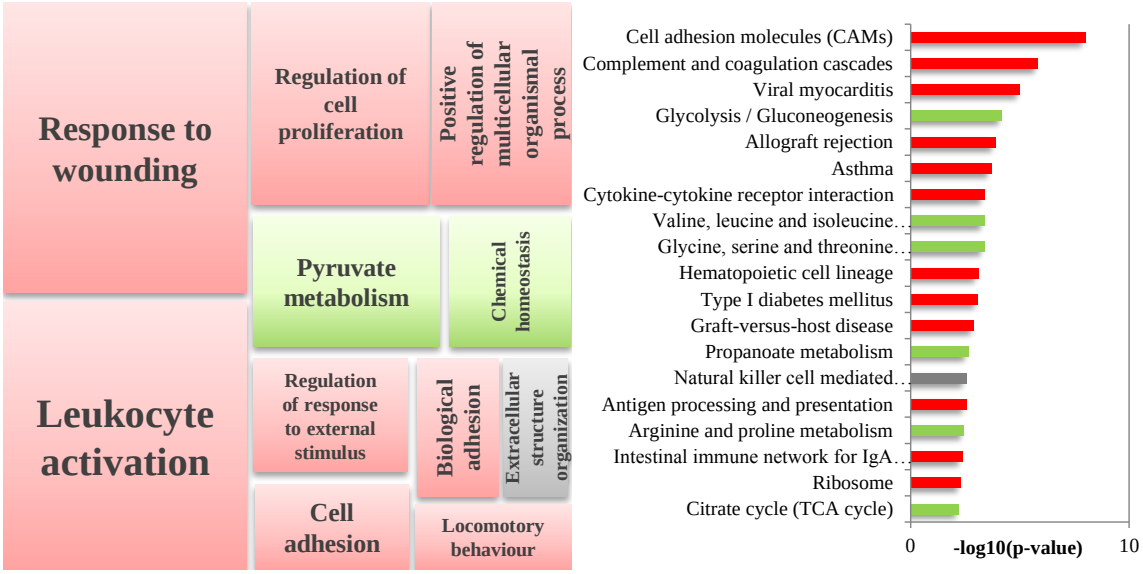


Figure 3.6: *(a)* GO term enrichment of biological processes for the up and down regulated genes, respectively, using representation adapted from REVIGO. *(b)* Enriched pathways - bar length is representative of the p-value, the bigger the bar the smaller the p-value; being the limits 9.65E-9 $\leq$ p-value <= 0.005. Red color, in this representations, represent up-regulated, green down regulated biological processes/pathways, when this type of analysis is made separately for up and down regulated protein-coding genes; grey color represents enriched GO terms/pathways that only appear when the combined analysis of up and down regulated protein-coding genes.

When looking separately at the up and down regulated genes it becomes clear that up regulated protein-coding genes GO term enrichment are related with immune response, regulation of cell proliferation and regulation of programmed cell death among others, response to wounding, biological adhesion and transnational elongation in ccRCC; whereas the down regulated ones are mostly associated with the carboxilic acid catabolism, oxidation-

reduction processes and fatty acid metabolism, coenzime metabolism and cellular amide metabolism among others.

In terms of enriched pathways, the up regulated protein-coding genes are part of up regulated pathways like cytokine-cytokine receptor interaction, complement and coagulation cascades, cell adhesion molecules (CAMs) ,natural killer cell mediated cytotoxicity, p53 signalling pathway and Jak-STAT signalling pathway, whereas the down regulated protein-coding genes of pathways like citrate cycle (TCA cycle), fatty acid metabolism, glycolysis / gluconeogenesis; pathways already associated with ccRCC (Zeng et al., 2014).

Only GO terms with false discovery rate (FDR) inferior to 0.05 were used to obtain the simplified representation of enriched terms in Figure 3.5 *(b)*. The boxes represent clusters of related GO terms and the size of the box reflects their p-value, the smaller the p-value, the bigger the box.

3.3.3 Weighted gene correlation network analysis

In order to proceed with an analysis that takes into account the relationship between the transcripts, independently of their differential expression, a weighted gene correlation network analysis through WGCNA package for R (Langfelder and Horvath, 2008) followed.

This type of analysis allows, through gene profile correlations and associations with specific traits, to find key genes for certain conditions, in this case ccRCC.

In the construction of the correlation network, a signed network was constructed in order to keep track of the correlation type between gene expression profiles and the use of the Spearman correlation in order to minimize the expression differences between lincRNA/potentially new lincRNA and protein-coding genes, since in this type of correlation the data is firstly ranked and then correlated, instead of using the Pearson linear correlation, where the values are correlated as are presented.

The traits that will be taken into account for further analysis are: sample type (normal or tumor samples); pathologic stage (tumor stage of different patients - stage I, II, III and IV); patient vital status; Setd2 mutation and expression hence its association to ccRCC and lastly but not less important, Rab38 oncogene up regulated expression (due to a read-through chimera related with SETD2 depletion - unpublished results from our group) and VHL gene down regulated expression in tumor samples, VHL being the gene that is the main catalyst in ccRCC occurrence (Figure 3.7 *(a)*).

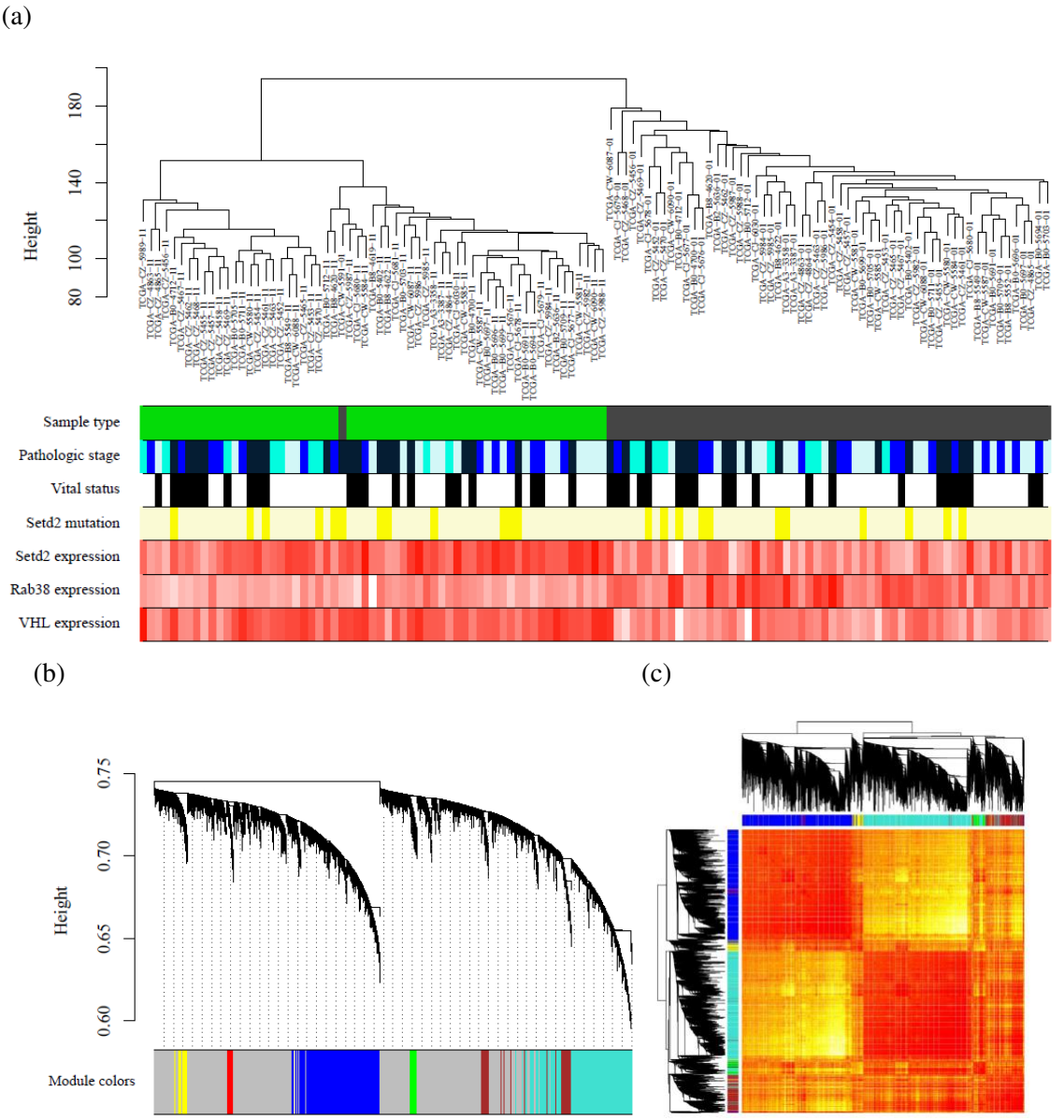


Figure 3.7: WGCNA network construction. *(a)* Sample dendogram and traits heatmap, with a hierarquical clustering of the samples based on the gene expression. For the heatmap interpretation, followa the color code explication: **Sample type** - green for normal samples, grey for tumor samples; **Pathologic stage** - clear tone of blue for tumor in stage I, darkest tone of blue for stage IV, intermediates tones for stage II and III; **Vital status** - white for alive, black for deceased; *Setd2 mutation* - clear yellow for not mutated, yellow for mutated; **Setd2**, **Rab38** and **VHL expression** - white for lower, red for high relative expression. *(b)* Gene dendogram and modules defined in the correlation network. *(c)* Hetmap of genes, present in the network defined modules, adjacency. Red color represents high positive adjacency; white values high negative correlation.

The module detection is based on hierarchical clustering of the expression data and allows to define clusters of highly interconnected genes, in the case of the signed network, genes that are positively correlated (Figure 3.7 *(b)*). The grey module color represent s genes that have no module assigned, whereas the others represent defined modules, in

this case six modules were assigned, being the turquoise and blue modules the ones with more genes. In Figure 3.7 *(c)* it is possible to observe the form that modules, and genes inside modules correlate with other genes from other modules; clearly observable that genes from the same module are positively correlated (red color), whereas genes from different modules are negatively correlated (white / yellowish color), as it happens for genes between blue and turquoise module.

If instead of a signed network, an unsigned network would have been constructed, no distinction could have been done between positively and negatively correlated genes, because in this case the correlation is an absolute value, and only a high or low correlation could exist between genes. For example if a gene expression correlation with another gene is -0.99, in the unsigned network the adjacency would be near 1 and they would be assigned to the same module; contrary, in the signed network they are assigned to different modules.

Module trait relationship

The next step in the network analysis is the data reduction towards the identification of modules highly correlated with traits. Each module has a defined *"module eigengene"* - ME - which can be considered as representative of the gene expression profiles in the module. When the module eigengenes of the defined modules are correlated with the traits, we obtain the modules that are more correlated to the traits, in this case, the modules that showed an absolute correlation bigger than 0.5, as well as p-value inferior to 0.5, towards at least one of the interest traits are the interest modules to study - blue, brown and turquoise(Figure 3.8 *(a)*). The genes present in the blue modules are genes that are up regulated in the tumor samples, whereas the genes from the brown and turquoise modules represent genes that are down regulated, but not mandatory defined as differentialy expressed by *cuffdiff* tool.

The protein-coding genes from the blue modules are genes are enriched in genes related with positive regulation of immune system process, defense response and cell activation; turquoise module protein-coding genes with monovalent inorganic cation transport, excretion, cation homeostasis and acetil-CoA metabolism; lastly, brown module protein-coding genes are related with carboxilic acid catabolism, cation transport and amine biosynthesis.

When looking at the proportion of protein-coding, lincRNA and potentially new lincRNA (Figure 3.8 *(b)*) it becomes clear that plenty lincRNA/potentially new lincRNA are correlated with protein-coding genes and plausibly have an important function in ccRCC.

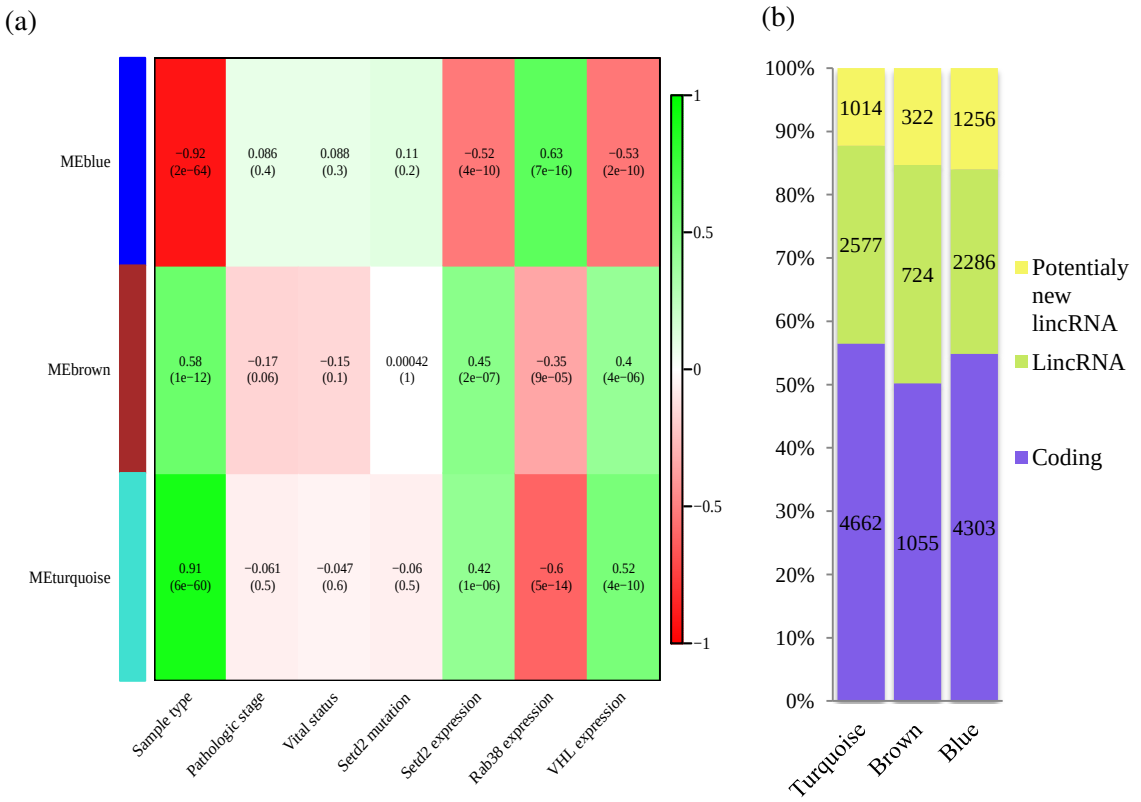


Figure 3.8: WGCNA modules analysis. *(a)* Module-trait relationship depicting correlations and associated Fisher asymptotic p-value between brackets for blue, turquoise and brown module, with several traits. *(b)* Gene proportion and associated number of different gene groups present in the interest modules.

In order to find interest genes in this modules and reduce the gene universe in the modules, genes were filtered based on gene membership measure - correlation of the gene expression pattern with the module eigengene; the ones with module membership higher than 0.8 were further filtered based on their gene significance towards the traits, being this gene significance the correlation between the gene expression profile and the trait, being the absolute value of gene significance superior than 0.5.

As observable in Figure 3.8, the turquoise and blue module are highly correlated with the sample type (absolute correlation value superior to 0.90 and small p-values). This modules are an optimal start point in order to uncover lincRNAs in the remaining module genes, highly associated with normal/ tumor sample. Curiously, genes already associated with ccRCC, SETD2 and VHL, show the same type of correlation with the same modules, whereas RAB38 has an opposite correlation with them.

If clustering the first 200 genes more correlated with sample type from blue and turquiose module, is possible to perfectly separate the normal samples from the tumor samples (Figure 3.9 A and B). When closing-up on the gene cluster, is possible to observe a co-expression/correlation of protein-coding genes, involved in biological functions enriched in ccRCC, and lincRNAs, although the last ones have a relative expression inferior

to the protein-coding genes.

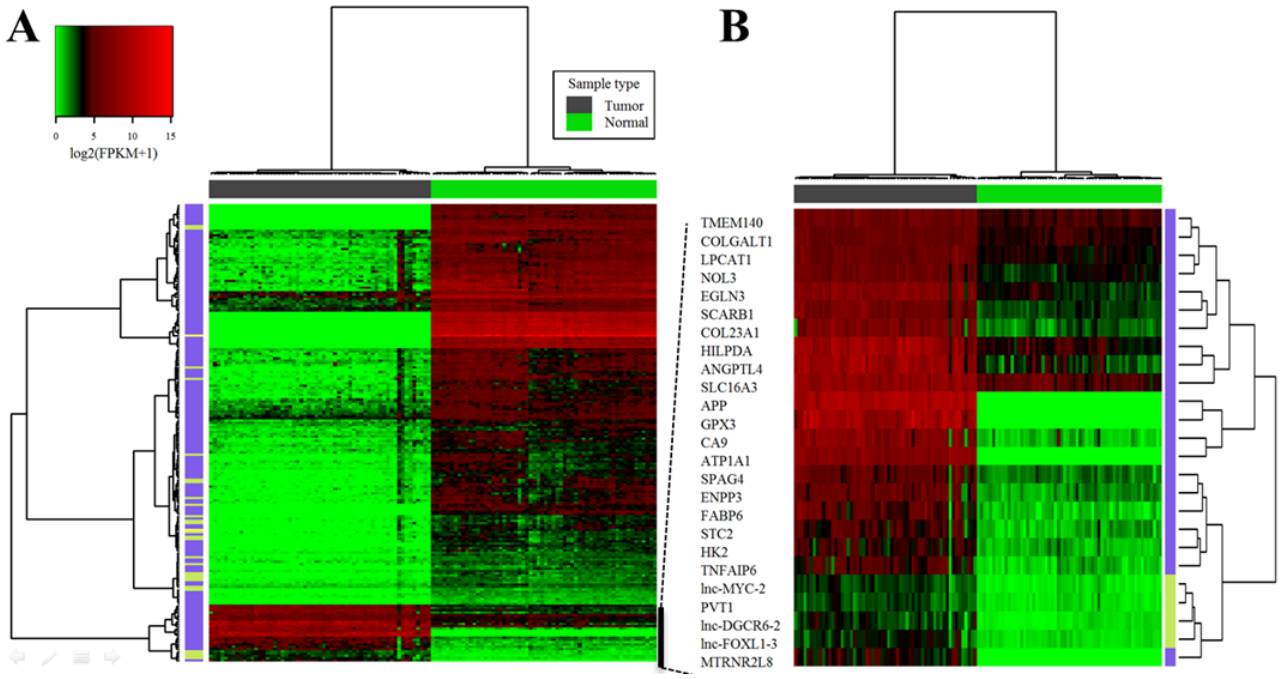


Figure 3.9: **A** Heatmap of the top 200 genes more correlated with sample type (171 protein-coding genes; 27 lincRNAs and 2 potentially new lincRNAs). **B** Close-up in gene clustering in order to show co-expression/correlation between protein coding-genes and lincRNAs. Row colors represent gene type: blue - protein-coding genes, green - lincRNA, yellow - potentially new lincRNA.

One of the lincRNA up regulated and present in this list is PVT1; that has been already associated with cancer. In colorectal cancer it generates antiapoptotic activity and an abnormal expression of PVT1 was a prognostic indicator for patients with this cancer type (Takahashi et al., 2014); by transfecting colorectal cells with PVT1-small interference RNA, Takahashi et al. (2014) showed that the cell had a significant lost their invasion and proliferation, with TGF-β pathway and apoptotic signals significantly activated.

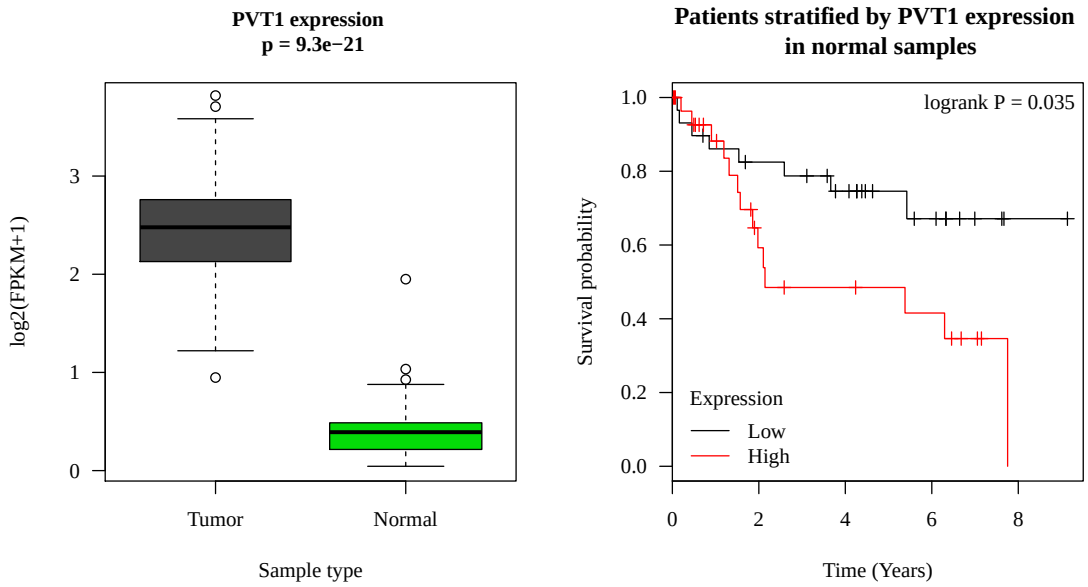
PVT1 expression levels in tumor tissue were significantly higher that the ones in the normal samples (Figure 3.10 (s)). Hence having matched normal/tumor samples, and if looking at the PVT1 expression in normal samples (patients were divided in two groups based on the median PVT1 expression in normal samples - if superior **high** else wise **low expression**), ccRCC patients normal samples with high PVT-1 expression had a significantly poorer prognosis when compared to low PVT1 expression levels.

Recent discoveries (Tseng et al., 2014) find that PVT1 can control the levels MYC (up regulated expression levels in ccRCC - Figure 3.10 (b)) by stabilizing the MYC protein, leading the authors believe that PVT1 interferes with MYC normal phosphorylation of threonine 58 induced degradation stabilizing the proteinn and increasing its levels.

MYC is a cell cycle regulator whose expression is induced by HIF (Rini et al., 2009) and it has been suggested that MYC pathway activation is essential in ccRCC develop-

ment and progression. Several pathways related to ccRCC (metabolism, signal transduction, translation, etc.) contain genes that are MYC targets, like BCL2, CCND1, PCNA, PGK1 and VEGFA (Rydzanicz et al., 2013); being all this genes present in the same blue module as PVT1, as well as MYC gene itself.

(a)



(b)

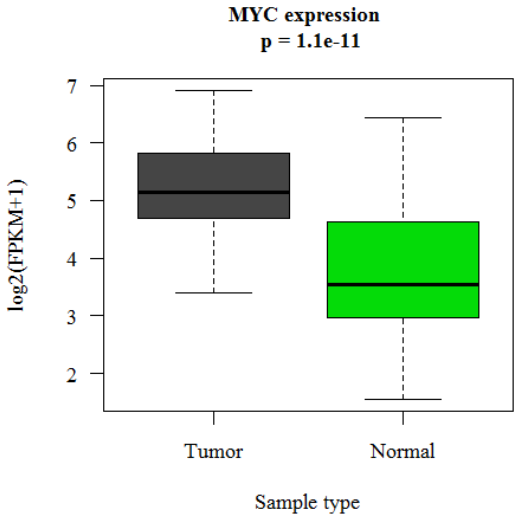


Figure 3.10: (a) PVT1 lincRNA expression in tumor and normal samples and Kaplan–Meier overall survival curves for 60 patients with ccRCC classified according to PVT1 expression level in normal samples. (b) MYC gene expression levels in tumor and normal samples.

This over expression of PVT1 in ccRCC combined with the literature may suggest a putative role for PVT1 in ccRCC progression, further studies being needed as well as

more data integration in order to ensure his plausible role in ccRCC.

With this type of overall analysis is possible to identify differentially expressed genes (2029 differentially expressed genes in ccRCC), including protein-coding genes, lincRNAs as well as potentially new lincRNA (found 5549 potentially new genes). On the other hand, using the weighted gene correlation network analysis allows to find genes highly correlated to certain traits, for example PVT1 possible involvement in ccRCC.

Chapter 4

Final Remarks and Future Perspectives

In this work, one of the most challenging part was the part corresponding to the human lincRNA catalogue construction. Several databases have information corresponding to lincRNA annotation, but not much consensus exist between them, fact that made us construct a human catalogue with a unified version of all lincRNAs across them. This catalogue allowed to achieve a final number of 38135 lincRNAs.

The different data formats in different databases, the instability of the databases that keep changing the gene identification and available data formats, brought another level of complications never expected before.

Due to lack of data integration, the merge of the different information present in the databases into one unique catalogue was a struggling process in order to define the best parameters for that to happen. This factor lead us to compromise and acknowledge our catalogue imperfections. Thus, 1.56% of the annotations present in the catalogue derive from the merge of two different annotated lincRNA from the same database due to an exon overlap with an annotated lincRNA from a different database.

This kind of problem may let us think which annotation is correct... from database A or from database B? but it actually makes me ask: ”What is in fact a gene and what defines an exon?”. I cannot answer to that question, but I rather suggest that a gene is a portion of DNA transcribed into RNA in a certain moment, controlled by several other factors. In order to respond to it, to have better annotations out there, more data integration has to occur as well a global effort for that to happen.

The second part presented in this thesis, and the main objective, was to characterize the lincRNAs expression profile in ccRCC. The RNA-seq dataset was large, comprising 124 samples (62 tumor normal paired samples), increasing the difficulty of the project. The first step was to create the transcriptome composition of ccRCC, using all this samples, in order to find potentially new lincRNAs. 5549 new transcripts were identified and were here designated as potentially new lincRNAs, since the possibility of being new pseudogenes could not be discarded. Thus, further analysis, would be needed to clearly classify all the new transcripts.

The second step was to asses differential gene expression, where a total of 2129 genes were significantly altered between tumor and normal samples (including protein-coding genes, annotated and potentially new lincRNAs). Not a very large number of lincRNAs were found to be differentialy expressed, in part because of the statistical tool used (Cufflinks). Due to their low expression levels, for many of the lincRNAs the statistical test was not even effectuated. Thus, we believe that to study low expressed genes, such as lincRNAs, other statistical tools (like edgeR or DESeq R packages) should be more appropriate.

To complement our analysis we used a weighted gene correlation network analysis, in order to take into account the relation between transcripts co-expression. This type of analysis revealed very promising results, allowing to find lincRNAs highly correlated with protein-coding genes and also associated to several cancer traits. Further data mining with this results has to be made, integrating it with other type of available data and other available bioinformatic tools.

In a summary, I would say that this analysis allowed to give the first steps in order to understand the lincRNAs importance and leaves an open window into unravelling new clues for ccRCC occurrence and aggressiveness. Further approaches, computational analysis and also laboratory experiments should be preformed to exhaustively understand the mechanisms involved in ccRCC.

Bibliography

American Cancer Society, 2014. Cancer Facts & Figures. *Atlanta: American Cancet Society*, .

Atak, Z. K., Gianfelici, V., Hulselmans, G., De Keersmaecker, K., Devasia, A. G., Geerdens, E., Mentens, N., Chiaretti, S., Durinck, K., Uyttebroeck, A., *et al.*, 2013. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS genetics*, **9**(12):e1003997.

Audenet, F., Yates, D. R., Cancel-Tassin, G., Cussenot, O., and Rouprêt, M., 2012. Genetic pathways involved in carcinogenesis of clear cell renal cell carcinoma: genomics towards personalized medicine. *BJU international*, **109**(12):1864–70.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, **25**(18):1915–27.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., *et al.*, 2005. The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, **309**:1559–1563.

Cech, T. and Steitz, J., 2014. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell*, **157**(1):77–94.

Cheetham, S. W., Gruhl, F., Mattick, J. S., and Dinger, M. E., 2013. Long noncoding RNAs and the genetics of cancer.

Chen, L.-L. and Carmichael, G. G., 2010. Long noncoding RNAs in mammalian cells: what, where, and why? *Wiley interdisciplinary reviews. RNA*, **1**:2–21.

Cheng, W., Zhang, Z., and Wang, J., 2013. Long noncoding RNAs: new players in prostate cancer. *Cancer letters*, **339**(1):8–14.

Chu, Y. and Corey, D. R., 2012. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, **22**(4):271–4.

Cock, P. J. a., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, **38**(6):1767–71.

Cohen, R. J., 2014. Pathology of Clear Cell Renal Cell Carcinoma.

Crumley, S. M., Divatia, M., Truong, L., Shen, S., Ayala, A. G., and Ro, J. Y., 2013. Renal cell carcinoma: Evolving and emerging subtypes. *World journal of clinical cases*, **1**(9):262–275.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., *et al.*, 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, **22**(9):1775–89.

Djebali, S., Davis, C. a., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.*, 2012. Landscape of transcription in human cells. *Nature*, **489**(7414):101–8.

Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y., and Segal, E., 2012. Genomics: ENCODE explained. *Nature*, **489**(7414):52–5.

Gingeras, T. R., 2007. Origin of phenotypes: genes and transcripts. *Genome research*, **17**(6):682–90.

Gupta, A. K. and Gupta, U. D., 2014. *Animal Biotechnology*. Elsevier.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., *et al.*, 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**(7235):223–7.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L., *et al.*, 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, **477**(7364):295–300.

Guttman, M. and Rinn, J. L., 2012. Modular regulatory principles of large non-coding RNAs. *Nature*, **482**(7385):339–46.

Huang, D. W., Sherman, B. T., and Lempicki, R. A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**(1):1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, **4**(1):44–57.

Huarte, M. and Rinn, J. L., 2010. Large non-coding RNAs: missing links in cancer? *Human molecular genetics*, **19**(R2):R152–61.

Illumina Inc., 2013. An Introduction to Next-Generation Sequencing Technology. .

Jonasch, E., Futreal, P. A., Davis, I. J., Bailey, S. T., Kim, W. Y., Brugarolas, J., Giaccia, A. J., Kurban, G., Pause, A., Frydman, J., *et al.*, 2012. State of the science: an update on renal cell carcinoma. *Molecular cancer research : MCR*, **10**(7):859–80.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., *et al.*, 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(28):11667–72.

Langfelder, P. and Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, **9**(1):559.

Larkin, J., Goh, X. Y., Vetter, M., Pickering, L., and Swanton, C., 2012. Epigenetic regulation in RCC: opportunities for therapeutic intervention? *Nature reviews. Urology*, **9**(3):147–55.

Linehan, W., Srinivasan, R., and Schmidt, L., 2010. The genetic basis of kidney cancer: a metabolic disease. *Nature Reviews Urology*, **7**(5):277–85.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M., 2012. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, **2012**:251364.

Ljungberg, B., Cowan, N. C., Hanbury, D. C., Hora, M., Kuczyk, M. a., Merseburger, A. S., Patard, J.-J., Mulders, P. F. a., and Sinescu, I. C., 2010. EAU guidelines on renal cell carcinoma: the 2010 update. *European urology*, **58**(3):398–406.

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., and Stahl, F., 2013. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, **24**(1):22–30.

Nagalakshmi, U., Waern, K., and Snyder, M., 2010. RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, **Chapter 4**(January):Unit 4.11.1–13.

Pinthus, J. H., Whelan, K. F., Gallino, D., Lu, J.-P., and Rothschild, N., 2011. Metabolic features of clear-cell renal cell carcinoma: mechanisms and clinical implications. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, **5**(4):274–82.

Qian, X., Ba, Y., Zhuang, Q., and Zhong, G., 2014. RNA-Seq technology and its application in fish transcriptomics. *Omics : a journal of integrative biology*, **18**(2):98–110.

Qiu, M.-T., Hu, J.-W., Yin, R., and Xu, L., 2013. Long noncoding RNA: an emerging paradigm of cancer research. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, **34**(2):613–20.

Qu, Z. and Adelson, D. L., 2012. Identification and comparative analysis of ncRNAs in human, mouse and zebrafish indicate a conserved role in regulation of genes expressed in brain. *PloS one*, **7**(12):e52275.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, **13**:341.

Rezende, R. B., Drachenberg, C. B., Kumar, D., Blanchaert, R., Ord, R. A., Ioffe, O. B., and Papadimitriou, J. C., 1999. Differential diagnosis between monomorphic clear cell adenocarcinoma of salivary glands and renal (clear) cell carcinoma.

Rini, B. I., Campbell, S. C., and Escudier, B., 2009. Renal cell carcinoma. *Lancet*, **373**(9669):1119–32.

Rosner, I., Bratslavsky, G., Pinto, P. A., and Linehan, W. M., 2010. The clinical implications of the genetics of renal cell carcinoma. *Urologic oncology*, **27**(2):131–6.

Rydzanicz, M., Wrzesiński, T., Bluyssen, H. a. R., and Wesoły, J., 2013. Genomics and epigenomics of clear cell renal cell carcinoma: recent developments and potential applications. *Cancer letters*, **341**(2):111–26.

Sanger, F., Nicklen, S., and Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**(12):5463–5467.

Shi, X., Sun, M., Liu, H., Yao, Y., and Song, Y., 2013. Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer letters*, **339**(2):159–66.

Strachan, T. and Read, A., 2010. *Human molecular genetics*. Garland Science, 4 edition.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, **6**(7):e21800.

Takahashi, Y., Sawada, G., Kurashige, J., Uchi, R., Matsumura, T., Ueo, H., Takano, Y., Eguchi, H., Sudo, T., Sugimachi, K., *et al.*, 2014. Amplification of PVT-1 is involved in

poor prognosis via apoptosis inhibition in colorectal cancers. *British journal of cancer*, **110**(1):164–71.

The Cancer Genome Atlas, 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**(7456):43–9.

The ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, **306**(5696):636–40.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L., *et al.*, 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, **7**(3):562–78.

Tsai, M.-C., Spitale, R. C., and Chang, H. Y., 2011. Long intergenic noncoding RNAs: new links in cancer progression. *Cancer research*, **71**(1):3–7.

Tseng, Y.-Y., Moriarity, B. S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P., Reuland, B., Guenther, K., Beadnell, T. C., *et al.*, 2014. PVT1 dependence in cancer with MYC copy-number increase. *Nature*, **512**(7512):82–86.

Ulitsky, I. and Bartel, D. P., 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**(1):26–46.

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J., and Mestdagh, P., 2013. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic acids research*, **41**(Database issue):D246–51.

Wain, H. M., Bruford, E. a., Lovering, R. C., Lush, M. J., Wright, M. W., and Povey, S., 2002. Guidelines for human gene nomenclature. *Genomics*, **79**(4):464–70.

Wang, Z., Gerstein, M., and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1):57–63.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y., *et al.*, 2014. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic acids research*, **42**(Database issue):D98–103.

Yadav, N. K., Shukla, P., Omer, A., Pareek, S., and Singh, R. K., 2014. Next Generation Sequencing: Potential and Application in Drug Discovery. *TheScientificWorldJournal*, **2014**:802437.

Zeng, Z., Que, T., Zhang, J., and Hu, Y., 2014. A study exploring critical pathways in clear cell renal cell carcinoma. *Experimental and therapeutic medicine*, **7**(1):121–130.

Appendix

LincRNA profile in ccRCC

A.1.1 TCGA patient and sample ID

Table A.1.1: TCGA patients and sample ID

Patient ID	Normal sample ID	Tumor sample ID	Patient ID	Normal sample ID	Tumor sample ID
TCGA-A3-3358	TCGA-A3-3358-11	TCGA-A3-3358-01	***TCGA-CW-5587***	TCGA-CW-5587-11	TCGA-CW-5587-01
TCGA-A3-3387	TCGA-A3-3387-11	TCGA-A3-3387-01	***TCGA-CW-5591***	TCGA-CW-5591-11	TCGA-CW-5591-01
TCGA-B0-4700	TCGA-B0-4700-11	TCGA-B0-4700-01	***TCGA-CW-6087***	TCGA-CW-6087-11	TCGA-CW-6087-01
TCGA-B0-4712	TCGA-B0-4712-11	TCGA-B0-4712-01	***TCGA-CW-6088***	TCGA-CW-6088-11	TCGA-CW-6088-01
TCGA-B0-5402	TCGA-B0-5402-11	TCGA-B0-5402-01	***TCGA-CW-6090***	TCGA-CW-6090-11	TCGA-CW-6090-01
TCGA-B0-5691	TCGA-B0-5691-11	TCGA-B0-5691-01	***TCGA-CZ-4863***	TCGA-CZ-4863-11	TCGA-CZ-4863-01
TCGA-B0-5694	TCGA-B0-5694-11	TCGA-B0-5694-01	***TCGA-CZ-4864***	TCGA-CZ-4864-11	TCGA-CZ-4864-01
TCGA-B0-5696	TCGA-B0-5696-11	TCGA-B0-5696-01	***TCGA-CZ-4865***	TCGA-CZ-4865-11	TCGA-CZ-4865-01
TCGA-B0-5697	TCGA-B0-5697-11	TCGA-B0-5697-01	***TCGA-CZ-5452***	TCGA-CZ-5452-11	TCGA-CZ-5452-01
TCGA-B0-5699	TCGA-B0-5699-11	TCGA-B0-5699-01	***TCGA-CZ-5453***	TCGA-CZ-5453-11	TCGA-CZ-5453-01
TCGA-B0-5703	TCGA-B0-5703-11	TCGA-B0-5703-01	***TCGA-CZ-5454***	TCGA-CZ-5454-11	TCGA-CZ-5454-01
TCGA-B0-5705	TCGA-B0-5705-11	TCGA-B0-5705-01	***TCGA-CZ-5455***	TCGA-CZ-5455-11	TCGA-CZ-5455-01
TCGA-B0-5709	TCGA-B0-5709-11	TCGA-B0-5709-01	***TCGA-CZ-5456***	TCGA-CZ-5456-11	TCGA-CZ-5456-01
TCGA-B0-5711	TCGA-B0-5711-11	TCGA-B0-5711-01	***TCGA-CZ-5457***	TCGA-CZ-5457-11	TCGA-CZ-5457-01
TCGA-B0-5712	TCGA-B0-5712-11	TCGA-B0-5712-01	***TCGA-CZ-5458***	TCGA-CZ-5458-11	TCGA-CZ-5458-01
TCGA-B2-5636	TCGA-B2-5636-11	TCGA-B2-5636-01	***TCGA-CZ-5461***	TCGA-CZ-5461-11	TCGA-CZ-5461-01
TCGA-B8-4619	TCGA-B8-4619-11	TCGA-B8-4619-01	***TCGA-CZ-5462***	TCGA-CZ-5462-11	TCGA-CZ-5462-01
TCGA-B8-4620	TCGA-B8-4620-11	TCGA-B8-4620-01	***TCGA-CZ-5463***	TCGA-CZ-5463-11	TCGA-CZ-5463-01
TCGA-B8-4622	TCGA-B8-4622-11	TCGA-B8-4622-01	***TCGA-CZ-5465***	TCGA-CZ-5465-11	TCGA-CZ-5465-01
TCGA-B8-5549	TCGA-B8-5549-11	TCGA-B8-5549-01	***TCGA-CZ-5467***	TCGA-CZ-5467-11	TCGA-CZ-5467-01
TCGA-B8-5552	TCGA-B8-5552-11	TCGA-B8-5552-01	***TCGA-CZ-5468***	TCGA-CZ-5468-11	TCGA-CZ-5468-01
TCGA-CJ-5676	TCGA-CJ-5676-11	TCGA-CJ-5676-01	***TCGA-CZ-5469***	TCGA-CZ-5469-11	TCGA-CZ-5469-01
TCGA-CJ-5677	TCGA-CJ-5677-11	TCGA-CJ-5677-01	***TCGA-CZ-5470***	TCGA-CZ-5470-11	TCGA-CZ-5470-01
TCGA-CJ-5678	TCGA-CJ-5678-11	TCGA-CJ-5678-01	***TCGA-CZ-5982***	TCGA-CZ-5982-11	TCGA-CZ-5982-01
TCGA-CJ-5679	TCGA-CJ-5679-11	TCGA-CJ-5679-01	***TCGA-CZ-5984***	TCGA-CZ-5984-11	TCGA-CZ-5984-01
TCGA-CJ-5680	TCGA-CJ-5680-11	TCGA-CJ-5680-01	***TCGA-CZ-5985***	TCGA-CZ-5985-11	TCGA-CZ-5985-01
TCGA-CJ-5681	TCGA-CJ-5681-11	TCGA-CJ-5681-01	***TCGA-CZ-5986***	TCGA-CZ-5986-11	TCGA-CZ-5986-01
TCGA-CJ-6030	TCGA-CJ-6030-11	TCGA-CJ-6030-01	***TCGA-CZ-5987***	TCGA-CZ-5987-11	TCGA-CZ-5987-01
TCGA-CW-5580	TCGA-CW-5580-11	TCGA-CW-5580-01	***TCGA-CZ-5988***	TCGA-CZ-5988-11	TCGA-CZ-5988-01
TCGA-CW-5581	TCGA-CW-5581-11	TCGA-CW-5581-01	***TCGA-CZ-5989***	TCGA-CZ-5989-11	TCGA-CZ-5989-01
TCGA-CW-5584	TCGA-CW-5584-11	TCGA-CW-5584-01	***TCGA-CZ-5989***	TCGA-CZ-5989-11	TCGA-CZ-5989-01